Prediction of the clinical outcome of NMIBC using Artificial Intelligence

Emanuele Frassini

Delft University of Technology

TU Delft

Prediction of the clinical outcome of NMIBC using Artificial Intelligence

by

Emanuele Frassini

Student name:	Emanuele Frassini
Student number:	5449995
Master specialization:	Applied Mathematics
Specialization track:	Computational Science and Engineering
Faculty:	EEMCS, Delft University of Technology
Thesis committee:	Prof.dr.ir. M.B. (Martin) van Gijzen
	F.K. (Farbod) Khoraminia
	Dr. A.B.T. (Alethea) Barbaro

to be defended publicly on Monday July 31^{st}, 2023 at 10:00 AM.
An electronic version of this dissertation is available at `http://repository.tudelft.nl/`.

TU Delft

Contents

Preface

As I come to the culmination of my master thesis project, I am overwhelmed with gratitude and emotion. I find it essential to acknowledge the individuals whose support and encouragement have made this journey possible.

First and foremost, I extend my heartfelt appreciation to my family, whose unconditional support and belief in my abilities have been a constant source of strength throughout my academic pursuits. I am immensely grateful to my friends from Delft, who have been more than just companions on this academic journey. Their willingness to share knowledge have enriched my learning experience and made the challenges easier to overcome. Your presence and support in both everyday life and study-sharing have been invaluable, and I cherish the memories we have created together. To my friends from Milan, even though you were physically distant, your presence was felt every step of the way. Your encouragement and belief in me have been a constant motivation, and I am grateful for the enduring connections we share. My heartfelt thanks go to my girlfriend, whose unwavering support, care, and understanding have been a guiding light in these months. Your continuous support and presence during the difficult moments have given me the strength to persevere and overcome hurdles.

I extend my deepest gratitude to Professor Martin van Gijzen for his remarkable mentorship and the invaluable exchange of ideas throughout this project. Your guidance and continuous help at every stage of this journey have been instrumental in shaping the direction of my research. I am equally thankful to my daily supervisor, Farbod Khoraminia, whose tireless efforts and expertise have been pivotal in the making of this project. Your valuable insights and support during the daily grind have been instrumental in bringing this thesis to fruition.

Finally, I would like to express my appreciation to all those who have played a part in my growth and development during this academic pursuit. Your support, both big and small, has left an indelible mark on my journey, and I am deeply grateful for each contribution. This master thesis project would not have been possible without the collective support of these incredible individuals.

Emanuele Frassini
Delft, July 2023

Nomenclature

Abbreviations

Abbreviation	Definition
NFM	Neuro Fuzzy Modelling
RF	Random Forest
GB	Gradient Boosting
SVM	Support Vector Machine
DT	Decision Tree
ET	Extreme Trees

1

Introduction

Bladder cancer (BC) holds the tenth position in terms of incidence rate among all cancers worldwide [1]. The bladder is a hollow organ in the lower abdomen in which urine is stored. This organ is made of many layers, including urothelium, lamina propria, muscle and fat tissue, from the innermost to the outermost. Cancer cells mostly initiate from the urothelium. Bladder cancer is usually classified into two main categories: non-muscle-invasive bladder cancer (NMIBC) and muscle-invasive bladder cancer (MIBC). Among the new diagnosed cases, approximatively 75% belong to the NMIBC group [2]. The term non-muscle invasive bladder encompasses tumors that are limited to the urothelium (such as carcinoma in situ or Ta) or those that invade the lamina propria (T1). Progression to MIBC involves the cancer's inclination to grow and invade the deeper layers of the bladder wall, whereas recurrence refers to the return of cancerous cells in the bladder following initial treatment. The general recurrence rate for NMIBC is between 60% to 70%, and the overall progression rate ranges from 20% to 30%. [6]. EAU guidelines stratify patients into low, intermediate, high and very high risk of progression to advanced disease. These guidelines are based on the current risk assessment tools such as EORTC and CUETO which rely on clinical and histopathological markers to classify patients for progression or recurrence [51]. The primary treatment approach for intermediate and high-risk patients is the use of Bacillus Calmette-Guerin (BCG) immunotherapy instillation administered locally [3].The latter is a therapy where the weakened tuberculosis bacteria is directly instilled into the bladder to stimulate an immune response that targets and destroys cancer cells. However 50% to 70% of patients will benefit from BCG treatment [4]. Patients who will not benefit from BCG immunotherapy or have high-risk disease may require radical cystectomy, a procedure aimed at removing the entire bladder, followed by chemotherapy and radiation [50]. Furthermore, patients who will progress to MIBC have a 5-year survival rate that ranges from 63% to 15% [63].

Accurate risk stratification tools are essential in determining the appropriate initial treatment. However, the existing methods used to predict the clinical outcome are not very reliable. For intermediate-risk patients, the EORTC model was impeded by the variability within this risk group, leading to both underestimation and overestimation of the risk of disease recurrence. On the other hand, for high-risk patients, the EORTC model overestimated the risk of disease recurrence, both at 1 and 5 years. In contrast, the CUETO model demonstrated poor calibration for disease recurrence, with an underestimation of the risk for low-risk patients and an overestimation for high-risk patients [7]. Statistical methods like the Cox proportional hazards (CPH) model and nomogram are used to predict bladder cancer outcomes using clinicopathological data. Nevertheless, recent advancements in artificial intelligence (AI) have shown superior accuracy in predicting disease recurrence and progression compared to statistical approaches [9, 39].

In this project, we try to overcome the problems of the current risk stratification models and help clinicians in the decision making. We aim at predicting the clinical outcome of NMIBC patients by using Artificial intelligence techniques, analysing clinicopathological data and histopathological images. The clinical outcome is divided in three binary classes: progression, high-grade recurrence and response to BCG treatment. The response to BCG treatment is to be considered a failure if the tumour progresses

to MIBC, or there is recurrence of high grade tumout until 6 months after completing BCG maintenance or presence of CIS tumour within 12 months after completing adequate BCG [8]. The two research questions that we want to address are:

- *Which AI methods can predict the clinical outcome of NMIBC patients by analyzing clinicopathological data?*
- *Which AI methods can predict the clinical outcome of NMIBC patients by performing an integrated analyses of clinicopathological data and histopathological images?*

We implemented six different classifiers. In particular, a neural network called adaptive neuro-fuzzy inference system, and five machine learning algorithms, such as random forest or gradient boosting. In the first stage, the clinicopathological data of the patients were used as input for the models. These include age, smoking status, grade and stage of the tumours. Secondly, the dataset was integrated with features related to the analysis of the cells nuclei in the histopatological images. To avoid a selection bias, no particular region of interests in the slides has been selected. The objective is to obtain an end-to-end unbiased method which can help clinicians in their decision making. Finally, the algorithms' performances were assessed with five common statistical metrics. The progonostic value of each feature was computed with a permutation importance analysis.

The report is organized as follows: Chapter 2 provides a systematic review about the state-of-the-art models developed so far in the field. Chapter 3 presents the mathematical background for the methods used in this study. The methods used and the research flow is presented in Chapter 4. The results obtained will be presented in Chapter 5. The discussion of the findings and the conclusions are found in Chapter 6. Chapter 7 focuses on potential improvements for the models. Tables with a a complete overview of all the models studied in the systematic review can be found in Appendix A. Appendix B presents the performances of all the models employed.

2

Systematic Review

This chapter is concerned with the review of the state-of-the-art methods that have been developed so far in the bladder cancer field. We performed a systematic review on five different databases, namely *Medline, Embase, Web of Science Core Collection, Cochrane Central Register of Controlled Trials* and *Google Scholar*. The keyword for the search were: *bladder cancer, machine learning* and *prediction*.The PRISMA chart of the inclusion criteria is depicted in Figure 2.1.

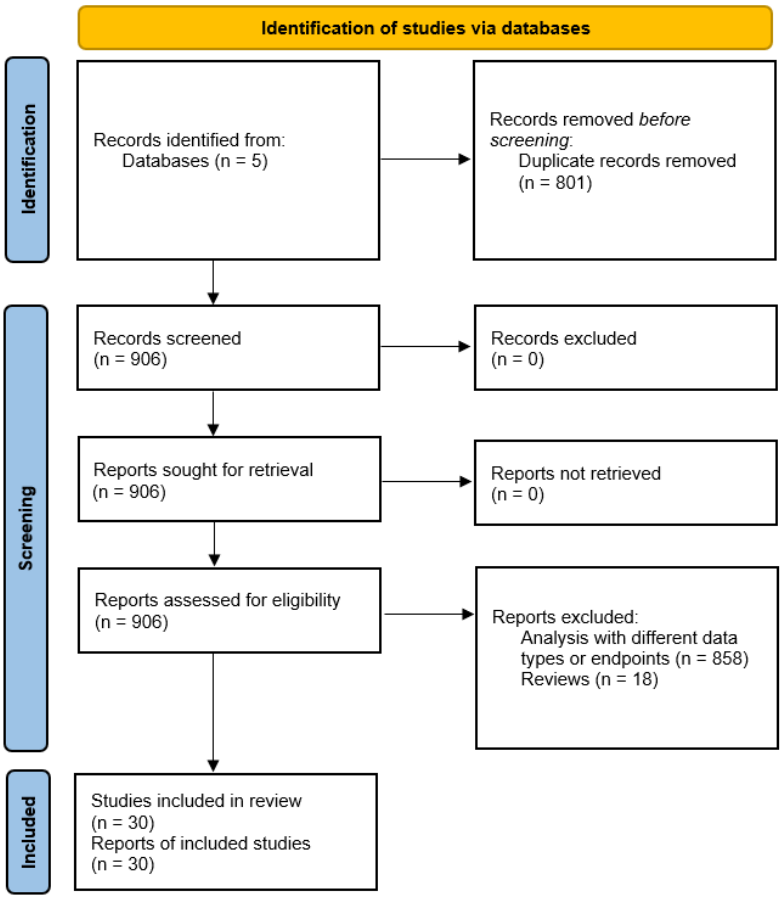

Figure 2.1: PRISMA chart of the systematic review

We included only papers that were in English, excluding conference abstracts and studies with different endpoints or type of data. As a result, thirty papers have been here reported. These studies focus on the

prediction of the clinical outcome of bladder cancer using artificial intelligence and statistical methods, analysing clinicopathological data. A summary analysis of all the papers can be found in Table A.2.

2.1. AI in Bladder Cancer prediction

Statistical algorithms, such as Cox proportional hazards (CPH) model and nomogram, have been used to predict the clinical outcome of bladder cancer using clinicopathological data [36]. The Cox proportional hazards model is a widely used statistical technique that can use clinical and pathological variables to predict the time to recurrence or death. CPH is used to estimate hazard ratios and confidence intervals to identify significant predictors of bladder cancer outcomes. The nomogram is a graphical calculating device that uses lines and scales to represent the relationships between multiple variables and predict a particular outcome. It is used to calculate an individualized risk score for patients, based on the values of selected predictors. This score can help guide clinical decision-making regarding the choice and intensity of therapy.

However, in recent times, artificial intelligence (AI) has exhibited better accuracy in predicting disease recurrence and progression than statistical methods [9, 39]. AI is a wide-ranging set of computational techniques that simulate human intelligence and has been extensively used in the medical field for computer-aided diagnosis (CAD) and computer-aided predictive (CAP) systems. Machine learning, which is one of the most important branches of AI, was developed to address problems in the medical domain [10].

Artificial neural networks (ANNs) are a type of AI techniques commonly employed in medical applications, including risk assessment for non-muscle invasive bladder cancer. ANNs are a subset of machine learning algorithms that mimic the structure and function of biological neural networks in the brain. ANNs are made up of interconnected nodes, which learn from patterns in data. During the learning process, the network is fed with labeled data, which helps to adjust the weights and biases of the nodes to minimize the error between the predicted and actual output. Once the ANNs are trained, they can predict clinical outcomes such as recurrence, progression of cancer patients using variables such as tumor size, grade, stage, or other clinical features. High accuracy can be achieved by training the network to predict the probability of an event, as a continuous output (regression) or as a discrete variable (classification). For instance, the network can predict the likelihood of disease progression or recurrence. The output of the ANNs can then be used to make informed clinical decisions.

Regression and classification are two supervised learning algorithms. In this subcategory of machine learning, the model is provided with labelled training data to learn the relationship between the input and output variables. In the context of bladder cancer diagnosis and outcome prediction, the vast majority of current machine learning applications falls under the category of supervised learning problems. These applications rely on labelled patient data to train the machine learning model to predict the clinical outcome of NMIBC based on a variety of clinical features. By leveraging supervised learning, machine learning models can help improve the accuracy and efficiency of NMIBC risk assessment and guide clinical decision-making [11].Supervised classification algorithms, for instance support vector machine (SVM), have been used to predict the clinical outcome of bladder cancer. SVM is a machine learning algorithm that is used to classify patients into different outcome groups based on their clinical and pathological characteristics. This algorithm works by finding the optimal hyperplane that maximizes the margin between the outcome groups, while minimizing the misclassification error. It has been used to predict the probability of bladder cancer recurrence to identify important predictors of bladder cancer outcomes.

Traditionally, the development of machine learning models necessitated the involvement of human experts who leveraged their domain knowledge to extract relevant features from raw input data. Following this feature extraction process, classification trees algorithms, such as Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB) and Extreme Trees (ET), were employed to discover mappings between the extracted features and the desired outputs [12]. Decision trees, for instance, are used to build a predictive model that uses a tree-like structure to classify patients into different outcome groups based on their clinical and pathological characteristics. The algorithm works by recursively splitting

the data into smaller and more homogeneous subgroups, based on the most significant predictors of clinical outcomes. The final tree structure provides a set of rules that can be used to predict a patient's outcome. Variable importance analysis is a technique commonly used with classification trees to identify the most significant features in predicting the target variable. This method is used to identify important predictors of bladder cancer outcomes, such as age, smoking habits or tumor characteristics. The trace-back process plays a key role in the development of an algorithm of clinical utility. Indeed, clinicians want to rely on a transparent model. Hence, ensuring an understanding of the learning process of the algorithm is of primary importance to effectively implement any model in a clinical setting.

2.2. Recurrence

Although the majority of patients with NMIBC have a good prognosis, a significant proportion will experience disease recurrence, which can have a significant impact on their quality of life. Various models have been developed to predict the risk of recurrence, including clinical, pathological, and molecular features. It is clinically known so far that the prior disease-recurrence rate and number of tumours are two of the most important prognostic factors for disease recurrence [45]. This section aims to review and compare the performance of different models for predicting recurrence in NMIBC, with a focus on their performance, feasibility, and clinical utility.

2.2.1. Statistical models

Xu et al. [38] aimed to predict the 2-year risk of recurrence of bladder cancer in a dataset of 71 patients. The study used age at the time of initial surgery, gender, histological grade, maximal tumor size in bladder lumen, tumor size, number of tumors and operation choice (transurethral resection of bladder tumor or radical cystectomy as features to build a nomogram. The features were selected using a Rad_Score model constructed with a support vector machine-based recursive feature elimination (SVM-RFE) approach and logistic regression. The nomogram achieved a moderate-to-high performance with a sensitivity of 0.778, specificity of 0.738, accuracy of 0.755, and an AUC of 0.822 in predicting recurrence. The study found that the most promising features extracted were image-related.

López de Maturana et al. [40] used age, gender, number of tumors, tumor stage and grade, tumor size, and treatment as features to predict time to first recurrence, defined as the reappearance of a NMIBC tumor following a previous negative follow-up cystoscopy. They employed a sequential threshold model on 1105 patients and achieved an AUC of 0.62. The study found that the role of common SNPs in predicting the risk of recurrence was limited and suggested that future studies should explore the integration of other genetic variants.

2.2.2. Artificial neural networks

The study conducted by Qureshi et al. [17] aimed to predict 6-month recurrence using a dataset of 212 patients and 14 features, including stage, grade, tumor size, tumor number, gender, EGFR status, smoking habit, histology, cis presence, metaplasia, architecture, site, c-erbB2, and p53 status. The authors employed an artificial neural network as a model for prediction, implemented with NeuralWorks Professional II/Plus software. The reported performance of the model, with a sensitivity of 0.7, specificity of 0.8, and accuracy of 0.75, suggests that it was able to identify patients at risk of recurrence with moderate accuracy. Despite the small number of patients for this analysis, it is important to mention that tumour size has shown to be the most significant feature in the prediction model.

The objective of Fujikawa et al.'s [42] investigation was to develop a prediction model for the recurrence of bladder cancer 15 years post-surgery, with a dataset comprising 90 patients. The analysis included a range of characteristics, such as tumor stage, grade, number, age, gender, tumor architecture, and mean nuclear volume estimates. The authors employed a Bayesian neural tool of SPSS Neural Connection 2.1 software as their model, which demonstrated a sensitivity of 0.33 and specificity of 0.4 in prognosticating recurrence. These findings suggest that the prediction model proposed by Fujikawa et al. had low efficacy in anticipating recurrence of bladder cancer after a period of 15 years. The low sensitivity and specificity values indicate an increased likelihood of false negatives and false positives, respectively. Several factors, including the limited size of the dataset and the selection of attributes

considered in the model, could have contributed to this low predictive power.

Buchner et al. [30] aimed to predict 5-year recurrence in 2111 patients using age, gender, tumor stage and grade (in transurethral resection of the bladder/TURB and RC), carcinoma in situ (TURB and RC), pathological lymph node status and lymphovascular invasion data. They employed an artificial neural network with a three-layer feed-forward perceptron architecture. Despite promising results in terms of specificity of 0.895 and accuracy of 0.74, the model sensitivity was low (0.4). The authors identified lymphovascular invasion, pathological T stage and pathological lymph node status as the most important features.

Catto et al. [39] aimed to predict the 80-month recurrence after surgery in 109 patients with bladder cancer. The study used a neuro fuzzy modeling approach and included several features such as stage, grade, age, sex, smoking exposure, and previous cancers to predict the recurrence. The results of the study showed that the neuro fuzzy model had a very high performance in predicting the recurrence with a sensitivity of 0.92, specificity of 0.9, accuracy of 0.92, and AUC of 0.98. The study also found that tumor grade, patient age, smoking history, and p53 expression were the most important features in predicting the recurrence. Overall, the study conducted by Catto et al. in [39] demonstrates the potential of using neuro fuzzy modeling to predict the post-operative recurrence of bladder cancer.

A second study, conducted by Catto et al. [41], focused on predicting the risk and timing of post-radical cystectomy tumor recurrence for 609 patients. They used gender, pathologic stage, pathologic grade, carcinoma in situ, and lymphovascular invasion as features and employed two neuro fuzzy modelling techniques combined in series to make predictions. The model achieved high performances, with a sensitivity of 0.81, specificity of 0.85 and a C-index of 0.92. The study found that tumor stage, lymphovascular invasion, and the number of removed lymph nodes were the most important features for predicting recurrence.

2.2.3. Ensemble learning

The study by Hasnain et al. [12] aimed to predict the recurrence of bladder cancer within 1, 3 and 5 years after radical cystectomy (RC) using operative findings at transurethral resection and radical cystectomy as well as pathology data. The dataset comprised 3499 patients who underwent RC. The study utilized a meta classifier based on support vector machine, bagged SVM, K-nearest neighbors, AdaBoost, random forest (RF) and gradient boosting trees algorithms, together with the concept of mutual information to uncover correlated parameters. The model, which has been trained and tested on 3499 patients, achieved a sensitivity of 0.739, a specificity of 0.714 and a low accuracy of 0.388 on 1 year recurrence. Considering a three-year time interval, the model achieved a good sensitivity and specificity, of 0.72 and 0.708 respectively, despite a low accuracy of 0.535. For the prediction of recurrence at 5 years, the model achieved a sensitivity of 0.7, specificity of 0.702 and accuracy of 0.588. The authors identified pathologic stage subgroup, pT stage, pN stage, pM stage, number of positive lymph nodes, pathologic positive lymph nodes, pathologic lymphovascular invasion, and clinical T stage (preoperative) as the most important features. The authors also highlighted the utility of the concept of mutual information in uncovering correlated parameters in the prediction of bladder cancer recurrence. Further studies may benefit from incorporating additional clinical and molecular features to improve the accuracy of the prediction model.

2.2.4. Combined models

Lucas et al. [36] compared three models, Cox proportional hazards (CPH), Boosted Cox model (BCM), and Random survival forest (RSF), to predict 1 and 5 years recurrence in 452 patients. WHO'73 grading, number of tumors as defined by the CUETO, the recurrence rate as defined by the EORTC and the age classification as defined by the CUETO have been given as input to the models. For the short time prediction, the performance varied slightly depending on the specific algorithm used, with CPH achieving the best performances (Se: 0.73, Sp: 0.59, Ac: 0.6 and AUC: 0.66). In the longer time interval of 5 years, the highest values were achieved by BCM, which achieved a sensitivity of 0.64, specificity of 0.61, accuracy of 0.60, and AUC of 0.72. It has been found that the EORTC and EAU risk classification systems showed slightly better predictive value than the CUETO risk stratification system in the

population under study. The subjectivity in assessing the histopathological variables and the difficulty in assessing grading and staging has been highlighted. The study also suggests that new prognostic markers are needed.

Dovey et al. [37] investigated the 1,2,5 and 10 years recurrence of bladder cancer. The study included 395 patients and aimed to predict recurrence using features such as multifocality, tumor stage, grade, and size. The model used in the study was a simplified version of the EORTC and CUETO models, and the performance was measured using the area under the curve (AUC). The model achieved an AUC of 0.7 for 1 year, 0.67 for 2 years, 0.69 and 0.66 for 5 and 10 years, respectively. The findings of the study suggest that the EORTC, CUETO, WHO '73 and WHO '04/16 models tend to underestimate recurrence. Additionally, the use of clinical covariates to predict recurrence may have reached its upper limits of accuracy, and studies have investigated molecular subtyping and genomic classification as an alternative. The authors identified several features that could be used to predict bladder cancer recurrence and developed a simplified model based on existing models. However, the model's performances were moderate, and it is suggested that there may be limitations to the use of clinical covariates in predicting recurrence accurately. Further research in the field, particularly into molecular subtyping and genomic classification, may offer alternative approaches to predict recurrence with higher accuracy.

2.2.5. Important features

NMIBC has a high recurrence rate, which poses significant challenges for clinicians and patients. Developing an accurate predictive model that identifies the features associated with recurrence is crucial for improving patient outcomes. In this section, we will examine the important features for a model aimed at predicting recurrence for NMIBC. We will review clinical and pathological features, molecular features, and demographic factors that have been shown to be associated with recurrence in NMIBC. By identifying the most critical features, it becomes possible to create a more accurate predictive model.

Pathological T stage has been found relevant in the outcome prediction by [12] and [30]. Also [17] mentions tumour size, one of the factors used to determine the T stage, as promising predictor. The pathological T stage for NMIBC (non-muscle invasive bladder cancer) refers to the extent of invasion of the tumor within the bladder lining, as determined by examination of the tissue under a microscope. The T stage is an important factor used in the TNM (tumor, node, metastasis) staging system to describe the pathological stage of bladder cancer, including NMIBC. In NMIBC, the cancer is limited to the inner lining of the bladder (urothelium) and has not invaded the muscle layer of the bladder wall. The pathological T stage for NMIBC is typically classified as Ta, T1, or CIS (carcinoma in situ), depending on the depth of invasion of the urothelium.

Lymphovascular invasion (LVI) has been determining in the models for [12] and [30]. LVI refers to the spread of cancer cells from the primary tumor into the lymphatic or blood vessels surrounding it. In the case of NMIBC, LVI can indicate a more aggressive tumor behavior and an increased risk of progression and recurrence. and is relatively uncommon but can occur in higher-grade tumors. When present, LVI is typically detected through microscopic examination of tissue samples taken during transurethral resection of the bladder tumor (TURBT).

Pathological lymph node status, which has been considered one of the key predictors in [12] and [30], refers to the presence or absence of cancer cells in the lymph nodes surrounding the bladder. In the TNM staging system, the lymph node status is used to describe the extent of the cancer and to guide treatment decisions. For NMIBC, the lymph node status is typically classified as N0 (no regional lymph node involvement) or N1-3 (involvement of one or more regional lymph nodes). It is important to note that high-grade NMIBC can spread to the lymph nodes and other distant sites, leading to a more advanced stage. The presence of lymph node involvement in NMIBC is typically determined through imaging studies or surgical removal and pathological examination of the lymph nodes.

Among the studies here included to predict recurrence, only [38] employed a nomogram attaining an AUC higher than 80% on 71 patients (Se: 0.778, Sp: 0.738 and Ac: 0.755). This finding still reflects the ability of such statistical technique to obtain moderate and high performance metrics when applied

to small data sets. Other statistical models, such as CPH, have been implemented with lower results ([16] attained sensitivity, specificity, c-index and accuracy around 60 %, with a slightly better AUC of 0.69). The performance of artificial intelligence techniques, such as artificial neural networks, varied among the studies analysed. For instance, the ANN with a three-layer feedforward perceptron architecture employed by [30] attained only 40% of sensitivity. This means that the model was not able to correctly identify a significant proportion of positive instances in the data set, even if the latter consisted of 2111 patients. Other types of neural networks, such as a Bayesian neural tool implemented by [42] poorly performed in this task, with sensitivity and specificity of 0.33 and 0.4, respectively. However, one specific kind of ANN appears to perform better on this specific task. The algorithm is called neuro fuzzy modelling and is a hybrid technique that combines the strengths of neural networks and fuzzy logic to create a more powerful and flexible modeling approach, useful in applications where there is a high degree of uncertainty and complexity in the data. [39] tried to predict recurrence in a cohort of 109 patients using this technique and obtained very high results in all the performance metrics (Se: 0.92, Sp: 0.9, Ac: 0.92 and AUC: 0.98). The same author implemented a similar model, combining in series two neuro-fuzzy modelling networks, in [41]. Sensitivity of 0.81, specificity of 0.85 and c-index of 0.92 have been achieved for a cohort of 609 patients, showing the potentiality of this technique even in a larger cohort.

2.3. Progression

Predictive models for the progression of NMIBC have been developed to aid in the clinical management of patients and to guide therapeutic decision-making. These models utilize various clinical and pathological factors, including tumor grade, stage, size, and presence of carcinoma in situ, to estimate the probability of disease progression. However, it is known to clinicians that the three most relevant risk factors for tumor progression are age, presence of multiple papillary tumours and a large tumour diameter ($>$ 3 cm) [45]. In recent years, there has been an increased interest in the development and validation of these models, as they have the potential to improve patient outcomes and optimize resource allocation in healthcare settings. However, there remains considerable variability in the performance and applicability of these models, and further research is needed to enhance their accuracy and utility in clinical practice.

2.3.1. Statistical model

Table A.19 shows the analysis conducted by López de Maturana et al. [40] on the time to first progression in patients diagnosed with non-muscle invasive bladder cancer. The dataset consisted of 1105 patients, and the features used for the analysis included area, age, number of tumors, tumor stage and grade, number of recurrences, and treatment. The authors used a sequential threshold model to analyze the data and predict the time to first progression, defined as the development of a muscle invasive tumor or a metastatic disease, or death because of UCB, after a previous diagnosis of NMIBC. This model is a type of regression analysis that involves adding variables to a model in a step-wise manner until the performance can no longer be improved. This approach allows for the selection of the most significant features and can improve the accuracy of the model. The performance of the algorithm was evaluated using the area under the curve (0.76), indicating that the model's predictive power was moderate. The findings of the study suggest that the role of common SNPs is limited in predicting the risk of recurrence in patients with NMIBC. The authors suggest that future studies should explore the integration of other genetic variants to improve the predictive performance of the model.

2.3.2. Artificial neural networks

Qureshi et al. [17] aimed to predict the progression of cancer within 6 months using different features including stage, grade, tumor size, tumor number, gender and EGFR status. The authors used an Artificial Neural Network (ANN) implemented with NeuralWorks Professional II/Plus software, which is a powerful tool that allows the user to customize the structure and parameters of the neural network to optimize its performance. The model achieved a sensitivity of 0.7, specificity of 0.82, and accuracy of 0.8, indicating the possibility of effectively predicting the progression of cancer within 6 months. However, it is important to note that the performance metrics used in this study do not provide a complete picture of the model's performance and the data set size (212) is relatively small. The findings of this

study suggest that EGFR status is the most important feature in predicting cancer progression within 6 months, which is clinically significant.

Table A.18 reports the study of Fujikawa et al.'s [42], which aimed to investigate the 15-year progression of NMIBC in a cohort of 90 patients. The study used tumor stage, grade, number of tumors, age, gender, tumor architecture, and estimates of mean nuclear volume as features and a Bayesian neural tool of SPSS Neural Connection 2.1 software as a model to predict the risk of tumor progression. The performance of the model was evaluated in terms of sensitivity and specificity and resulted in 1 and 0.67, respectively. The findings of the study revealed that patients judged to have a favorable prognosis using ANN analysis did not progress within the 15-year follow-up period.The choice of the model, Bayesian neural tool of SPSS Neural Connection 2.1 software, reflects the need to analyze the complex relationship between multiple features to predict tumor progression accurately. Bayesian neural networks have shown promising results as they can handle complex, noisy, and uncertain data. The extremely high sensitivity score indicates that the model correctly identified all patients who progressed, and the findings suggest that the model's performance was satisfactory in predicting tumor progression. However, the specificity score was low, indicating that the model had a high false-positive rate.

The study by Catto et al. [43] aimed to predict 80-month progression using a dataset of 117 patients. The features implemented in the algorithm were tumor stage, tumor grade, age, gender, smoking exposure, and previous cancers. An hybrid neural network that combined a fuzzy logic model trained on subgroups defined by hierarchical clustering algorithm has been developed. The model's performance was evaluated using sensitivity, specificity, accuracy, and Area Under the Curve, which were reported to be 0.88, 0.99, 0.94, and 0.99, respectively. The type of model selected is particularly useful when dealing with complex and heterogeneous datasets, as it allows for the identification of different subgroups of patients with similar clinical characteristics. The fuzzy logic model is used to predict the outcome based on the input features, while the hierarchical clustering algorithm is used to define the subgroups on which the fuzzy logic model is trained. This approach can lead to a more accurate and reliable prediction, as it takes into account the heterogeneity of the patient population. The very high performance achieved by the model, with an AUC of 0.99, suggests that the choice of the model was appropriate for the task at hand. The high sensitivity and specificity values indicate that the model has a high ability to correctly classify patients who will and will not experience progression. The small sample size of 117 patients, with a test set size of only 10%, may limit the generalizability of the findings and the results obtained. Overall, this study highlights the importance of carefully selecting the appropriate model and features when developing prediction models for bladder cancer progression.

Abbod et al. [44] aimed to predict the 100-month progression of bladder cancer in a dataset of 117 patients. The features used in the model were stage, grade, age, sex, smoking exposure and previous cancers. Two models have been implemented: a neuro-fuzzy model (NFM) and a multi-layered perceptron artificial neural network (ANN) with 15 hidden neurons, to predict the outcome. The NFM achieved sensitivity of 0.88, specificity of 0.99, accuracy of 0.94, while the ANN achieved sensitivity of 0.81, specificity of 0.95, accuracy of 0.89. The choice of the model used to predict the outcome of the disease is crucial for achieving high performance. The neuro-fuzzy model combines the advantages of both fuzzy logic and neural networks allowing for more precise and flexible modeling of complex systems. On the other hand, the multi-layered perceptron artificial neural network is a well-known and widely used model in the field of machine learning. However, the results showed that the neuro-fuzzy model outperformed the ANN in all the evaluation metrics. In terms of the findings, the study highlights smoking exposure as a significant risk factor for advanced bladder cancer disease.

2.3.3. Combined model

Dovey et al. [37] conducted a study with the aim of predicting bladder cancer progression over 1, 2, 5, and 10-year periods using a dataset of 395 patients. The study employed multifocality, tumor stage, grade, and size as features in the algorithm to predict outcomes. The authors developed a simplified model based on the EORTC and CUETO scoring systems by selecting only the most important features which most impacted on the outcome prediction. The model's performance was assessed using the Area Under the Curve (AUC), which was found to be 0.88 for 1 and 2-year progression, 0.84 for

5-year progression, and 0.82 for 10-year progression. Despite a decrease in AUC as the length of the time period increases, the model's performance remained consistently above 80%, indicating high discriminatory power to differentiate between progressing and non-progressing patients for each outcome. The study also found that the existing scoring systems, including EORTC and CUETO, tend to underestimate recurrence in bladder cancer patients. As a result, molecular subtyping and genomic classification are proposed as alternative approaches for future research. However, a limitation of the study is the small data set size and relatively large proportion of low-risk NMIBC patients, which may have introduced bias in the analysis of progression data.

2.3.4. Important features

Predicting the likelihood of progression from NMIBC is crucial for determining appropriate treatment strategies. Clinicopathological data, including patient age, tumor size, and grade, have been used to develop predictive models. In this section, we will try to identify the most important features for accurate predictions reported by the different models previously analysed. Understanding the key predictors of NMIBC progression can aid in the development of personalized treatment plans and improve overall patient care.

Tumour size and grade have been found to be determining for the outcome prediction in [37] and [43]. Tumor size refers to the physical dimensions of the cancerous tissue within the bladder. In NMIBC, tumor size is typically measured in centimeters and is often used to classify the cancer as either low-grade or high-grade. Low-grade NMIBC tumors are typically smaller in size and have a lower likelihood of progressing to muscle-invasive bladder cancer (MIBC), while high-grade tumors are larger and more aggressive, with a higher likelihood of progression. Tumor grade refers to the degree of abnormality of the cancer cells, which is determined by examining the cells under a microscope. In NMIBC, tumor grade is classified as either low-grade or high-grade. Low-grade NMIBC tumors have cells that closely resemble normal bladder cells, while high-grade tumors have cells that appear more abnormal and are more likely to grow and spread quickly. Tumor grade is an important factor in determining the appropriate treatment for NMIBC and predicting the likelihood of progression.

Smoking exposure has been relevant in the analysis conducted by [43] and [44]. Smoking exposes the bladder to carcinogenic compounds and increases the risk of DNA damage, which can lead to the development of bladder cancer. Patients with a history of smoking may require more frequent follow-up and surveillance to detect potential recurrence or progression of the cancer.

Additionally, epidermal growth factor receptor (EGFR) has been reported as key predictor in [17]. EGFR is a protein that is present on the surface of many types of cells, including the cells that line the bladder. This protein plays an important role in cell growth and division and its overexpression has been associated with the development and progression of many types of cancer, including non-muscle invasive bladder cancer. For NMIBC, EGFR expression has been found to be higher in high-grade tumors compared to low-grade tumors, suggesting a potential role in tumor aggressiveness and progression.

In order to predict the progression of NMIBC, a statistical method called sequential threshold model has been employed by [40], achieving an AUC of 0.76 on 1105 patients included in the study. The artificial neural network implemented by [17] on 212 patients showed good performance results, with sensitivity, specificity and accuracy of 0.7, 0.82 and 0.8, respectively.
However, neuro-fuzzy modelling proved to be the best performing algorithm also for this outcome, even with a small dataset size. [43] implemented a hybrid neural network that combines a fuzzy logic model trained on subgroups defined by hierarchical clustering algorithm, attaining very high AUC and specificity of 0.99 and high accuracy and sensitivity of 0.94 and 0.88, respectively. Furthermore, also [44] implemented neuro-fuzzy modelling on a cohort of 117 patients. A sensitivity of 0.88, specificity of 0.99 and accuracy of 0.94 proved the outstanding performances of this algorithm among the others.

2.4. Overview of the models

The prediction of recurrence and progression of NMIBC can be a challenging task. Among the studies analysed in this review, neuro-fuzzy modelling proved to be the best performing algorithm. The strength of this model lies in both the high performance achieved by the studies here presented and the interpretative power of the model. Therefore, our analysis will start with the implementation of such model. Furthermore, traditional machine learning techniques showed moderate results in the prediction of both outcomes. Ensemble trees structure in particular, have shown to provide a robust and accurate predictions together with a high interpretive power of the learning process. Hence, five traditional ML algorithms will be implemented, namely random forest, gradient boosting, support vector machine, decision trees and extreme trees. Subsequently, an analysis of the tumour slides will be performed. The aim of this analysis will be the extraction of important tumour-related features from the histopatologic images. The variables will be added to the existing predictors in our data set in order to enhance the reliability and accuracy of our forecasts.

3 Background

3.1. Data

In our study, both clinicopahtological data and histopatological images have been analyzed.

Clinicopathological data combine clinical and pathological information about the medical condition of a patient. For instance, patient symptoms, medical history, physical examination findings, laboratory test results, and pathological assessments are considered. Clinical data provide insights into the patient's status, such as age, gender, or smoking habits, and the clinical manifestation of the disease. Pathological data involve the analysis of tissues, cells, or biological samples to identify characteristics of the tumour.

Histopathological images are microscopic images of tissues or cells. Histopathology involves the preparation of thin slices of tissue, which are then stained using various techniques to enhance specific cellular components or highlight certain pathological features. The stained tissue sections are placed on glass slides and viewed under a microscope. Pathologists examine these type of images to identify and characterize various aspects, for example cellular morphology or tissue architecture.

3.2. Neuro-fuzzy modelling

Neuro-fuzzy modelling is a computational model that combines the strengths of artificial neural networks and fuzzy logic to create a hybrid intelligent system. This model integrates the learning capabilities of neural networks with the reasoning and linguistic expressiveness of fuzzy logic, by using a layered architecture of neurons.

3.2.1. Fuzzy logic

The fuzzy logic consists in a set of fuzzy rules. The rules are expressions of the form *if A then B* where *A* and *B* are labels of fuzzy sets determined by membership functions. In our context, the fuzzy sets involve only the first part. These are called Takagi and Sugeno's fuzzy rules [52]. For instance, the rules can be of the form *if velocity is high, then force =* $k \times (velocity)^2$.These expressions can be valuable to employ a model in a clinical setting. In particular, the rules give insight into the way the algorithm predicts the target variable. By following this approach, a clinician would be able to have a prediction followed by a set of rules which can be further evaluated and validated. Transparency in the reasoning behind algorithmic predictions significantly enhances the value of the predictions. It enables clinicians to understand and follow the underlying thought process, thus instilling greater confidence in the predictions made. Transparency allows clinicians to carefully inspect and confirm the reasoning.

Membership functions offer a more comprehensive representation than the traditional indicator function for classical sets. In essence, these functions provide a means to quantify the degree or grade of membership of an element to a fuzzy set. By employing membership functions, we can assign a numerical value to indicate how strongly an element belongs to a particular fuzzy set, allowing for

a more nuanced and flexible approach in handling uncertainties and vagueness. In contrast to the binary nature of classical sets, membership functions enable us to express the varying degrees of membership associated with elements in a set, facilitating a more realistic and versatile modeling of complex systems.

3.2.2. Architecture

This neural network's architecture is made up of five layers as presented in a simplified version in Figure 3.1.

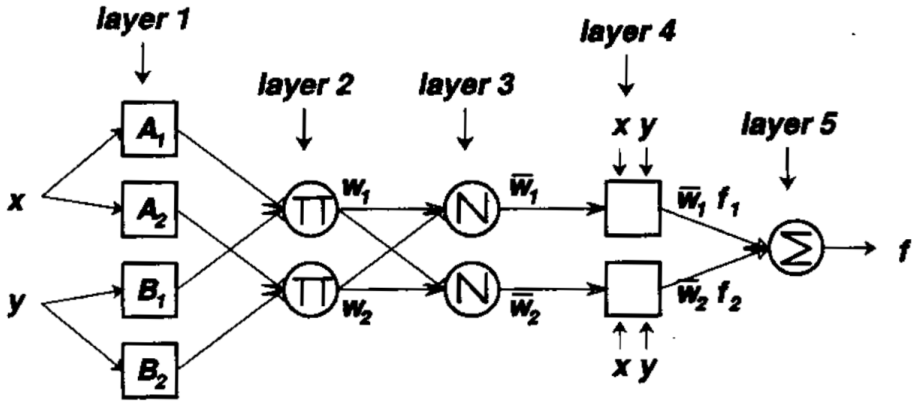

Figure 3.1: Simplified neuro-fuzzy modelling architecture with two rules [46]

The first layer takes the input data, which are standardized and fed to the model, and applies a Gaussian membership function (N_{ij}). The data consist of a n dimensional vector (x_j) for each sample, where n is the number of features. The output of the first layer will be $N_{ij}(x_j)$, where $i = 1, ..., m$ represent the rules. Layer 2 is made of a total of $m \cdot n$ nodes, where m is the desired number of fuzzy rules. Every node in this layer multiplies the inputs and sends the product to the next layer. To avoid overfitting of the network, an optimal value of $m = 2 \cdot n$ has been chosen. The product is of the form

$$w_i = \prod_{j=1}^{n} N_{ij}(x_i).$$

Each output represents the strength of the rule. In the third layer, the i^{th} node computes the ratio of the i^{th} rule's strength with respect to all the rules' strength. That is, the output of this layer is

$$\bar{w_i} = \frac{w_i}{\sum_{i=1}^{m} w_i}.$$

Layer 4 consist of m nodes, where each node has a function f_i. Here, the node function is multiplied by the strength computed in layer 3 ($\bar{w_i} f_i$) to be given as input for the last layer. The final layer, takes as input the sum over all the nodes of the rules and applies a sigmoid function to generate the final output [46]. Specifically, the output will be $\sigma(X)$, where

$$X = \sum_{i=1}^{m} \bar{w_i} f_i,$$

and

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

is the sigmoid function. Since our target variables are binary, if $\sigma(X) > 0.5$ we assigned label 1 to the input, and 0 otherwise.

3.3. ML models

3.3.1. Support vector machine

Support Vector Machine (SVM) is a prediction technique, rooted in statistical learning frameworks. With a given set of labeled training instances, an SVM training algorithm constructs a model that assigns new instances to one of two categories. It functions as a deterministic binary linear classifier. SVM maps the training examples onto points in space to maximize the separation between the two categories. Subsequently, new instances are projected onto the same space and classified based on which side of the separation they fall on [53].

Consider n data points of the form (x_i, y_i) $i = 1, ..., n$, where each x_i is a p-dimensional vector and y_i are either 1 or -1, indicating the class to which the point belongs. The algorithm aims at finding an hyperplane that maximises the distance between the points. A generic hyperplane equation is given by

$$w^T x - b = 0,$$

where w is the normal vector to the hyperplane. Assuming that the data is linearly separable, it is possible to choose two hyperplanes to separate the two classes of data. These hyperplanes are described by the following equations, for the positive and negative class respectively:

$$w^T x - b = 1,$$

$$w^T x - b = -1.$$

Since the distance between these planes is given by $d = \frac{2}{||w||}$, the goal is to minimize $||w||$. In order to avoid points in the margin between the two classes, the following costraints are imposed:

$$y_i(w^T x_i - b) \geq 1 \; \forall i = 1, ..., n.$$

Hence, the objective of this method is to solve an optimization problem:

$$min \; ||w||_2^2,$$

$$s.t. \; y_i(w^T x_i - b) \geq 1 \; \forall i = 1, ..., n.$$

If the data are not linearly separable, SVM applies a hinge loss function, given by

$$L_i = max(0, 1 - y_i(w^T x_i - b)).$$

The new function to minimize will then be

$$\lambda ||w||^2 + \frac{1}{n} \sum_{i=1}^{n} L_i,$$

for the parameter $\lambda > 0$.

3.3.2. Decision Tree

A decision tree is a classification method represented as a recursive division of the instance space [54]. The tree structure consists of nodes forming a directed tree with a ”root” node having no incoming edges, and all other nodes having exactly one incoming edge. The internal nodes split the instance space into sub-spaces based on specific discrete functions of the input attribute values. Typically, each test at an internal node considers a single attribute, dividing the instance space according to its value. Each leaf node is associated with a class representing the most suitable target value. To classify instances, the decision tree is traversed from the root to a leaf based on the test outcomes along the path. Decision trees can be visualized geometrically as a collection of hyperplanes, each orthogonal to one of the axes, for numeric attributes. The tree complexity, which affects accuracy, is controlled by stopping criteria and pruning methods. Common metrics to measure tree complexity include the total number of nodes, leaves, tree depth, and number of attributes used.

Consider a vector $x_i \in \mathcal{R}^n \; i+1,...,l$ and a label vector $y \in \mathcal{R}^l$, a decision tree divides the feature space in a recursive manner, grouping together samples with identical or similar labels [55]. Suppose at a node m the data are represented by Q_m, with n_m samples. A candidate split $\theta = (j, t_m)$, which consists of a feature j and treshold t_m, partitions the dafa in a right and left subset, i.e. $Q_m^{right}(\theta)$ and $Q_m^{left}(\theta)$. Indeed we have

$$Q_m^{left}(\theta) = \{(x, y)|x_j \leq t_m\},$$
$$Q_m^{right}(\theta) = Q_m \setminus Q_m^{left}(\theta).$$

The quality of a possible split for a node m is given by the Gini impurity function.The Gini index is a measure of impurity that quantifies the differences between probability distributions of the target attribute's values [54]. The objective is to select the variable θ that minimises the impurity, i.e. find θ^* such that:

$$\theta^* = argmin_\theta G(Q_m, \theta),$$

where

$$G(Q_m, \theta) = \frac{n_m^{left}}{n_m} H(Q_m^{left}(\theta)) + \frac{n_m^{right}}{n_m} H(Q_m^{right}(\theta)),$$
$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk}) \; ,$$

and

$$p_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k).$$

We can see that p_{mk} is indeed the proportion of class k observations for the node m.

3.3.3. Ensemble Trees

The objective of ensemble methods is to enhance generalization and robustness compared to a single estimator by aggregating the predictions of multiple base estimators constructed using the same learning algorithm. These algorithm are divided into two main categories: bagging and boosting methods.

Bagging methods

In ensemble algorithms, bagging methods belong to a group of algorithms that create multiple instances of an estimator by using random subsets of the original training set. These individual predictions are then combined to generate the final prediction. The purpose of these methods is to decrease the variance of a base estimator, such as a decision tree, by introducing randomness into its construction process and forming an ensemble. Bagging methods offer a straightforward approach to improve over a single model without requiring modifications to the underlying base algorithm. They are particularly effective in reducing overfitting and are best suited for strong and complex models [56]. In this class, random forest and extreme trees are perturb-and-combine methods, particularly tailored for trees. This approach involves creating a diverse set of classifiers by introducing randomness during the construction of each classifier. The ensemble's prediction is obtained by averaging the predictions made by the individual classifiers (Figure 3.2).

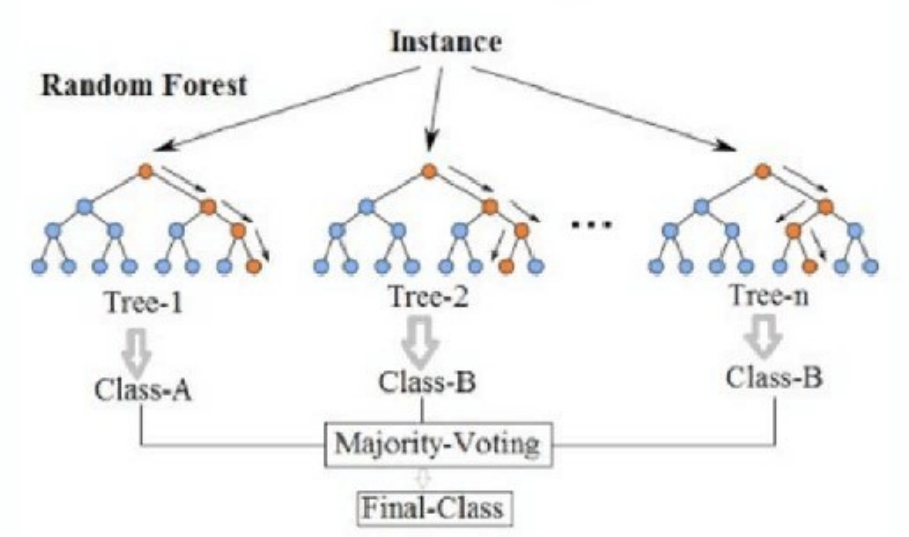

Figure 3.2: Random forest algorithm [62]

Random forests utilize a specific approach where each tree in the ensemble is constructed from a sample drawn with replacement (bootstrapped sample) from the training set [56]. Additionally, during the process of splitting each node while building a tree, the best split is determined either from all input features or from a random subset of size max_features. The main purpose of introducing these two sources of randomness is to reduce the variance of the forest estimator. Individual decision trees typically have high variance and are prone to overfitting. By injecting randomness into the construction of the forests, the decision trees tend to have somewhat independent prediction errors. When averaging the predictions of these diverse trees, some errors cancel out, leading to a reduction in variance. While this may slightly increase bias, the overall model performs better due to the significant reduction in variance, making random forests a practical and effective modeling technique.

In extreme trees, the level of randomness in split computation is taken a step further. Similar to random forests, a random subset of candidate features is considered. Instead of seeking the most discriminative thresholds, these are randomly generated for each candidate feature. From these randomly generated thresholds, the best one is chosen as the splitting rule. This approach typically results in a slightly greater increase in bias. However, it allows for a bit more reduction in the model's variance. In essence, the increased level of randomness in extremely randomized trees helps further decrease the model's variability, making it a useful trade-off for improved overall performance.

Boosting methods

Boosting is an ensemble meta-algorithm primarily used to reduce bias and also variance in supervised learning. It belongs to a family of machine learning algorithms that aim to transform weak learners into strong ones [57]. In this context, a weak learner is defined as a classifier that shows only a slight correlation with the true classification. It can perform better than random guessing when labeling examples. On the other hand, a strong learner is a classifier that demonstrates a high correlation with the true classification, performing exceedingly well in making accurate predictions. Boosting assigns weights to the outputs of each individual tree, with higher weights given to incorrect classifications from the first decision tree as input to the next tree. Through multiple iterations, this method combines the weak rules into a single, robust prediction rule with enhanced accuracy and power.

Consider a input data x_i and a prediction $\hat{y}_i$. The prediction will be of the form

$$\hat{y}_i = F_M(x_i) = \sum_{m=1}^{M} h_m(x_i),$$

where h_m are the weak learners estimators and M is the total number of estimators. Recursively, we

can compute

$$F_m(x) = F_{m-1}(x) + h_m(x).$$

To build a new fitted tree h_m, the aim is to minimize a sum of losses L_m on the previous ensemble F_{m-1}. Specifically,

$$h_m = argmin_h L_m = argmin_h \sum_{i=1}^{n} l(y_i, F_{m-1}(x_i) + h(x_i),$$

where $l(y_i, F(x_i))$ is the log-loss function. Since we are dealing with a classification problem, the mapping from the continuos value predicted by the trees to the class needs to be specified. For the log-loss function, the positive class ($y = 1$) is modelled as

$$p(y_i = 1|x_i) = \sigma(F_M(x_i)),$$

where σ is the sigmoid function.

3.3.4. Permutation importance

Permutation feature importance is a method used to inspect and understand a fitted estimator. It is particularly valuable for models to have insights of the training procedure. In our context, this technique is essential because a prediction without any insight into how the model reached its conclusions lacks clinical relevance. This technique measures the impact of each feature on the model's performance by randomly shuffling the values of a single feature and observing the decrease in the model's score [57]. When the relationship between the feature and the target is broken by the shuffling, the drop in the model's score indicates how much the model relies on that specific feature. One of the significant advantages of this technique is its model-agnostic nature, allowing it to be applied repeatedly with various permutations of the feature.

Given a predictive model m, a dataset D, and a number of repetitions K, the technique computes the reference accuracy s of the model. Subsequently, each feature j is randomly shuffled to produce a corrupted version of the dataset $\tilde{D}_{jk}$ for each repetition $k \in K$. The score s_{jk} is then computed again on $\tilde{D}_{jk}$. Finally the importance of the feature j is defined by

$$i_j = s - \frac{1}{K} \sum_{k=1}^{K} s_{jk}.$$

3.4. Image segmentation model: Stardist

Image segmentation is a widely employed method in digital image processing and analysis that involves dividing an image into several parts or regions, relying on the characteristics of the pixels within the image. Recent successful learning-based methods include segmenting cells at the pixel level and then grouping them, or localizing bounding boxes and refining shapes subsequently [47]. However, in crowded cell scenarios (Figure 3.3a), these approaches may encounter segmentation errors, such as merging neighboring cells inaccurately or suppressing valid cell instances due to bounding box limitations.

To overcome these problems, the model developed by [48] has been chosen. The approach consists of localizing cell nuclei using star-convex polygons, which offer a superior shape representation compared to bounding boxes and eliminate the need for shape refinement. Star-convex polygons are polygons that contain a point from which the entire boundary is visible. For this purpose, a convolutional neural network has been trained to predict a polygon for each pixel, representing the cell instance at that position. The distances $(r^k_{i,j}{}_{k=1}^{n})$ from each pixel indexed by (i, j) to the boundary of the object to which the pixel belongs are regressed along a set of n predefined radial directions with equidistant angles (Figure 3.3b). To make sure that the object contains pixels, only polygons with sufficient high object probability ($d_{i,j}$) are considered (Figure 3.3b). The object probability is computed as as the normalized Euclidean distance to the nearest background pixel. After obtaining polygon candidates along with their corresponding object probabilities, non-maximum suppression (NMS) is conducted to obtain the

ultimate set of polygons, where each polygon signifies an individual object instance [47] (Figure 3.3c). NMS is a computer vision technique used to choose a singular entity from a set of overlapping entities.

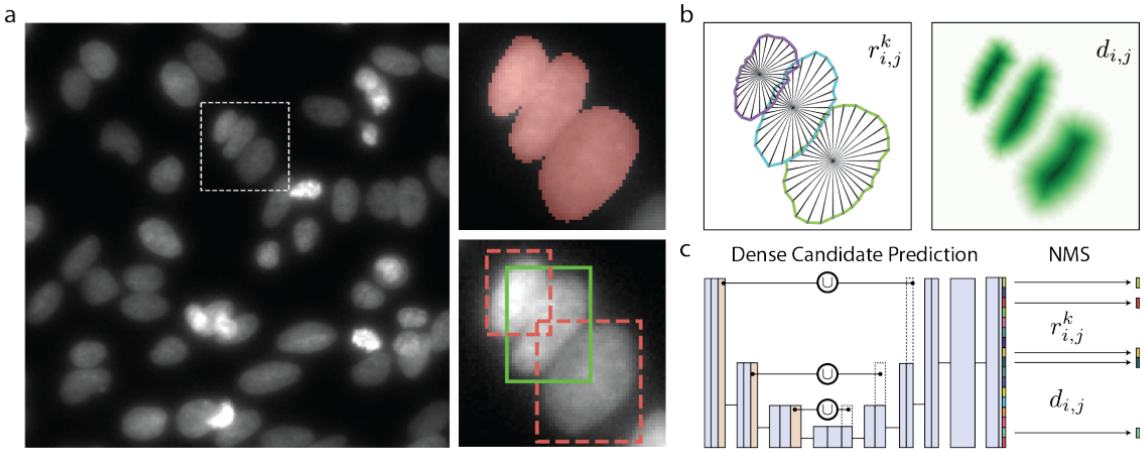

Figure 3.3: Stardist model architecture [47]

3.5. Clustering technique: FlowSOM

FlowSOM is a clustering technique commonly used in flow cytometry analysis. It was developed as an extension of Self-Organizing Maps (SOM), which is an unsupervised machine learning algorithm. FlowSOM is specifically designed to handle high-dimensional data, where each cell is characterized by multiple parameters. FlowSOM is often used to identify distinct cell populations.

The FlowSOM pipeline is depicted in Figure 3.4. After reading the data, the algorithm consists of three additional steps: the creation of a self organising map, the generation of a minimal spanning tree and the computation of a meta-clustering result [49].

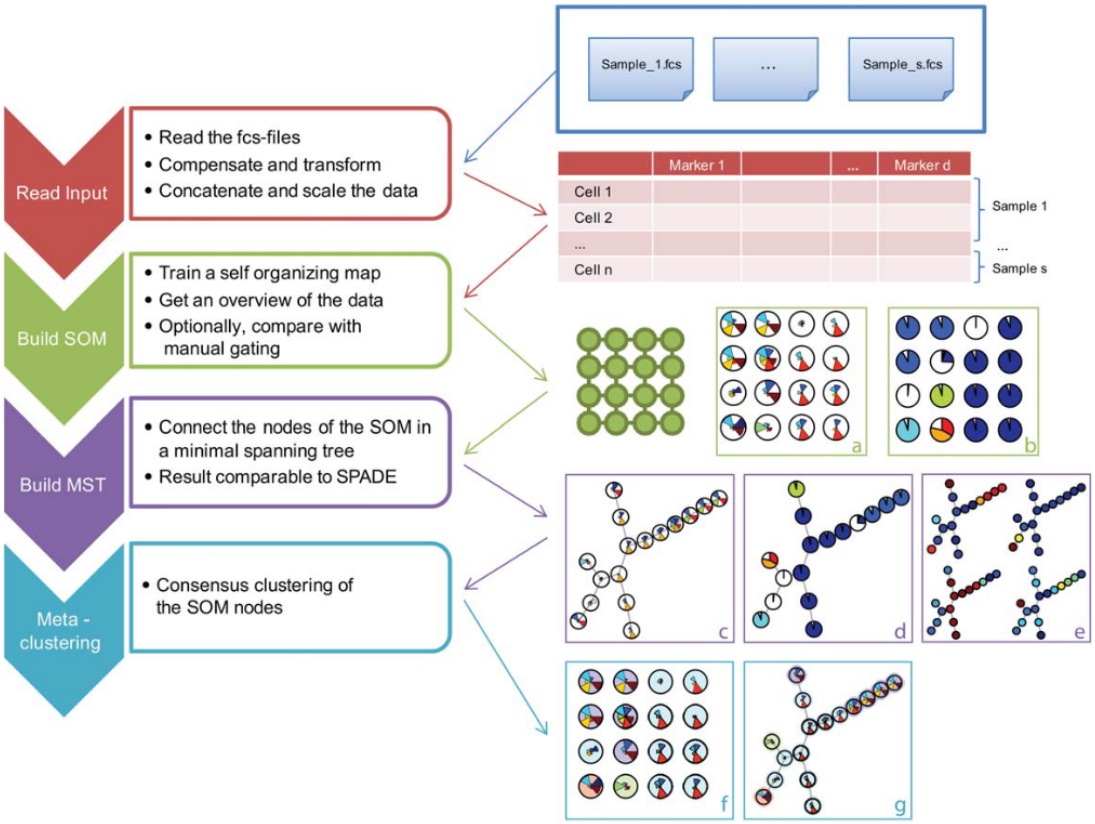

Figure 3.4: FlowSOM pipeline [49]

A self-organizing map (SOM) is an unsupervised method used for both clustering and dimensionality reduction. It involves training a specific type of artificial neural network on a discretized representation of the input space. Furthermore, a significantly larger number of clusters compared to the anticipated cell types are employed. This approach also reveals information about subpopulations that might have been overlooked during the initial manual gating process. The self organising maps algorithm starts with the creation of a grid of nodes, in which each node represents a point in the multidimensional input space. The grid is trained in a manner where nodes in close proximity exhibit greater resemblance to each other compared to nodes connected by a longer path [49]. The algorithm starts by randomly initiate the node weight vectors in the map, and randomly picking an input vector V from the nuclei dataset, with index v. Later, each node in the map is traversed and the euclidean distance between V and the map's node weight vector is computed. The node that has the smallest distance to V is called best matching unit (BMU), with index u. Thus, the weight vectors (W) of the nodes in the proximity of BMU are updated by pulling them closer to the input vector. Specifically, for each iteration s,

$$W_v(s+1) = W_v(s) + \theta(u, v, s)\alpha(s)(V - W_v(s)),$$

where α is the learning rate and and θ is a restraint due to the distance to the BMU.

The result of the self organising map can be analysed in a minimal spanning tree (MST). An MST is a subset of the edges in a connected, edge-weighted graph that links all the vertices together without forming any cycles and has the lowest total edge weight possible. As a result, the nodes of the SOM grid are connected to the ones they are the most similar to, taking the multidimensional topology of the data into account [49].

3.6. Metrics

In order to evaluate the performance of our predictive models, we have employed five evaluation metrics. These metrics include accuracy, F1 score, AUC, sensitivity, and specificity. Accuracy is a widely

used metric that measures the overall correctness of our predictions. It represents the proportion of correctly predicted instances out of the total number of instances in our dataset. A higher accuracy score indicates that our model has made more correct predictions, while a lower accuracy score suggests that our model's predictions are less reliable. The F1 score is a measure of the model's accuracy, particularly in cases where the dataset is imbalanced. It considers both precision and sensitivity to provide a more balanced evaluation. Precision calculates the ratio of correctly predicted positive instances to all instances predicted as positive, while recall measures the ratio of correctly predicted positive instances to the actual number of positive instances. The F1 score combines these two measures to give a comprehensive assessment of our model's performance. The AUC is a metric commonly used in binary classification tasks. It measures the model's ability to distinguish between positive and negative instances by plotting the true positive rate against the false positive rate. The AUC score ranges from 0 to 1, where a score closer to 1 indicates a model with better discriminative ability. Sensitivity measures the proportion of actual positive instances correctly identified by our model. It is particularly useful in scenarios where identifying positive instances is crucial. A high sensitivity score indicates that our model has a lower chance of missing positive instances. Specificity represents the proportion of actual negative instances correctly identified by our model. High specificity suggests that our model has a lower chance of incorrectly labeling negative instances as positive.

The presentation of the results is accomplished through the use of boxplots. A boxplot is a statistical visualization tool that provides a concise summary of the distribution of a dataset. It consists of a box that represents the interquartile range (IQR), which encapsulates the middle 50% of the data. Within the box, a line is drawn to indicate the median, which represents the central tendency of the dataset. Extending from the box, whiskers are drawn to capture the range of the data. The length of the whiskers is 1.5 times the IQR. Any data points falling beyond the whiskers are considered outliers and are plotted individually as distinct points, showed as black circles.

4

Methods

4.1. Experimental setup

In order to conduct the experiments, both clinicopathological data and histopatological images were available at Erasmus MC. Three different analysis have been performed. The first two consisted in the use of only clinicopathological data and only histopatological images, followed by an integrated study of both cinical and image-related features.

The flow diagram of the study is depicted in Figure 4.1. The initial phase of our process involved gathering clinicopathological data and histopathological images. To facilitate analysis, we divided the images into smaller sections known as patches and conducted segmentation to detect individual cell nuclei within these patches. Subsequently, a variety of features were extracted from the identified nuclei, and clustering techniques were applied to group them together, resulting in the creation of a dataset specifically focused on image-related features. In the subsequent stage, the image-derived dataset was integrated with the clinicopathological data, merging the information obtained from both sources. This merged dataset, containing a combination of image-related features and clinicopathological data, was then used as input for multiple classifiers. To gain insights and make predictions, we ran various algorithms on the integrated dataset. These classifiers employed different techniques to analyze the data and make predictions based on the available features. Additionally, we performed variable importance analysis, which helped identify the most influential features in the prediction process. By determining the relative importance of each variable, we gained valuable insights into which factors were most critical for accurate predictions.

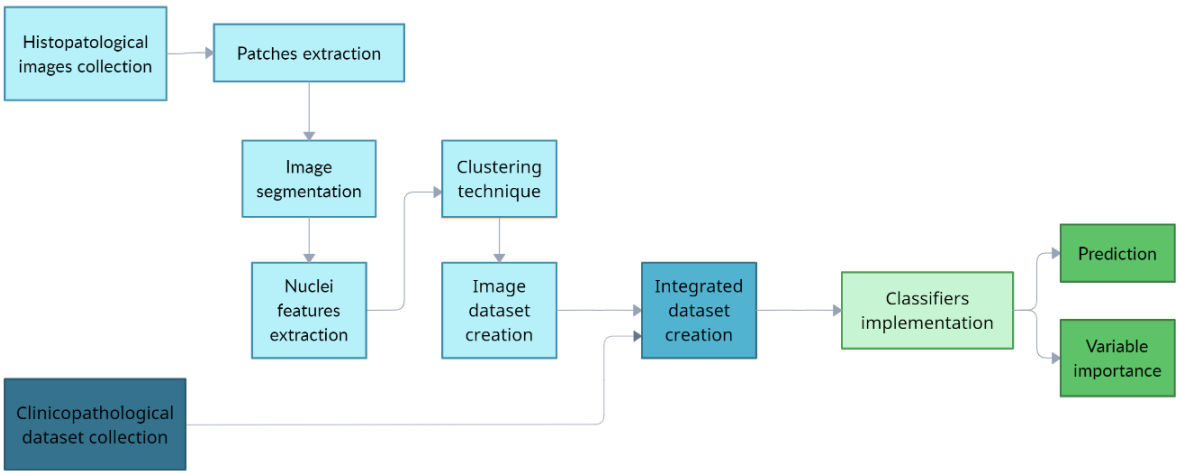

Figure 4.1: Flow diagram of the study

4.1.1. Clinicopathological data

A collection of tumor samples was gathered from patients with high risk non-muscle invasive bladder cancer who had undergone at least five or six initial treatments of BCG between 2000 and 2018. These samples were obtained from six different hospitals, including five in the Netherlands (Erasmus University Medical Center Rotterdam, Franciscus Gasthuis and Vlietland Rotterdam, Amphia Breda, Haga and Reinier de Graaf Gasthuis, Delft) and one hospital in Norway (Stavanger University Hospital). The inclusion of diverse patient cohorts enhances the generalizability and robustness of our findings.

The initial group comprised 1134 patients with High Risk Non-Muscle Invasive Bladder Cancer (HR-NMIBC). Patients who did not receive a sufficient number of BCG instillations and those who were diagnosed with Muscle Invasive Bladder Cancer (MIBC) during follow-up were excluded. The dataset encompassed patient information from diverse cohorts and was originally stored in SPSS, a widely used statistical software for data management and analysis. In order to consolidate the data from different sources, we merged the datasets into a single Excel file.

Nine features were derived from the clinicopathological data and used as predictors for the models. These include age, gender, smoking status, number and size of the tumours, presence of concomitant CIS, stage and grade of the tumours and history of cancer, according to previous studies' findings [9]-[45].

4.1.2. Histopatological images

For our study, the stained slices have been scanned using a whole-slide imaging scanner at high resolution (x80 magnifications). Image patches were extracted from whole slide images at 40X magnification level, with 512X512 pixel size with 25% overlap. The extraction process has been performed using the algorithm developed by [59]. This tissue segmentation algorithm divides the whole slide images in smaller patches, and classifies the patches based on the type of tissue. In order to perform our analysis, we automatically selected patches that contained urothelium tissue while excluding those that consisted of muscle, stroma, damaged tissue, or blood.

In order to maintain the impartiality of our image selection process, we excluded the WSI that had been punched. This measure was taken to eliminate any potential bias in our selection, ensuring that all chosen images remained unaltered and representative of the entire cohort. Additionally, by avoiding the inclusion of punched images, we aimed to prevent any chance of eliminating regions within the tissue that may contain the highest grade or stage information.

4.2. Segmentation model

In order to segment the image patches, we employed the Stardist algorithm, which is explained more in detail in Section 3.4. The images underwent initial preprocessing by converting them to black and white scale. This conversion is essential in improving the contrast between different structures and objects in the H&E stained images. As a result, segmentation becomes more precise as the varying intensities in different regions are more apparent. By simplifying the color complexity of the H&E stained image, the conversion to black and white reduces data dimensionality, consequently facilitating the identification and segmentation of specific structures or objects based on intensity variations.

The ground truth for 200 H&E stained images have been created using Qupath, an open source software for bioimage analysis. This step was essential as it provided us with a reliable benchmark against which we could measure the accuracy of the segmentation results. By establishing a ground truth, we were able to effectively determine the extent to which the segmentation algorithm correctly identified and delineated the desired areas within the images. With these ground truth at hand, we run the pretrained model named *2D_versatile_he*, developed by [49]. In a later stage, the convolutional neural network (CNN) was retrained over the course of 400 epochs, utilizing 160 images and labels for training purposes. The performance of the retrained model was then evaluated on the remaining 20% of the labeled images. The evaluation of segmentation performance is conducted based on various metrics, considering the Intersection over Union (IoU) threshold τ. This threshold determines the minimum

overlap required between the ground truth and predicted boxes for a prediction to be considered a true positive.

4.2.1. Nuclei features extraction

Once the images have been segmented, hundreds of millions of cell nuclei have been detected by the algorithm in the whole image cohort. Therefore, features related to the morphology of the nuclei have been extracted to gain information with prognostic value for our analysis. This process has been performed using a popular image processing library named scikit-image [60]. The images are represented as NumPy arrays, with 512 rows and 512 columns. Each value of the array represents a pixel in the original image.

The algorithm works by creating bounding box (blue lines in Figure 4.2), that include any detected object. Afterwards, information such as the area, perimeter, eccentricity or solidity of the object are computed and stored in a dataframe. Hence, we gave the images we segmented using Stardist to the model and a total number of 22 features have been computed for each detected nuclei. Additionally, we kept track of the position of the nuclei inside the patch, and of the position of the patch inside the whole slide image. In this way, we were able to trace back to the exact position of the indentified object in the original image.

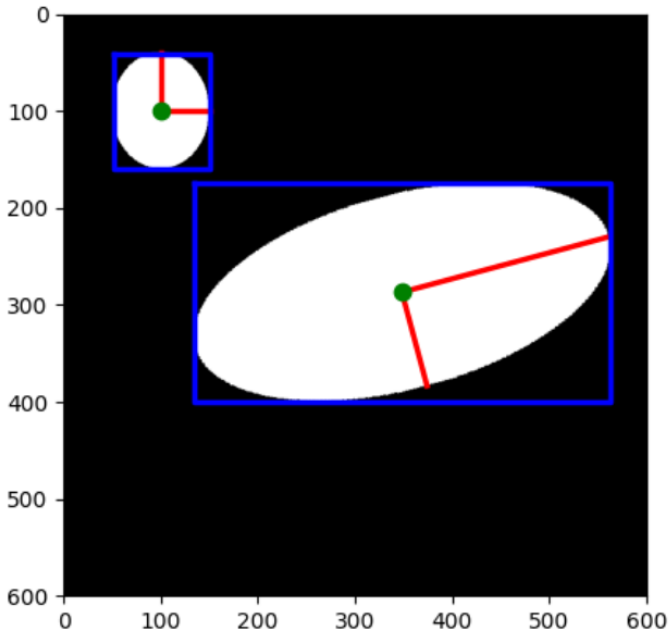

Figure 4.2: Example of features extraction
`https://scikit-image.org/docs/stable/_images/sphx_glr_plot_regionprops_001.png`

Nevertheless, the process of analyzing thousands of nuclei per patient presents a further challenge. The abundance of data renders it impractical to utilize all the nuclei for classifying a patient's clinical outcome. To address this issue, we devised various strategies aimed at selecting a representative subset of nuclei for each patient, enabling meaningful analysis. One approach involves computing the average value of the features across all the nuclei for each patient. However, this method is not advisable due to its sensitivity to outliers and the potential bias introduced by artifacts. The average values may not accurately represent the majority of cells, leading to skewed results and misinterpretations. To overcome these limitations, we turned to an advanced clustering technique that has demonstrated promising results in the field of biology. This methodology allows us to group nuclei based on their similarity, capturing underlying patterns and facilitating the identification of representative clusters. By selecting representative nuclei from these clusters, we created a more condensed and meaningful dataset for each patient. Employing clustering techniques enhances our ability to identify distinct subgroups of cells with shared characteristics, potentially related to specific disease subtypes or clinical

outcomes. This approach enables us to capture the heterogeneity within the patient cohort, while also providing a more manageable dataset for subsequent analysis.

4.3. Clustering technique

After segmenting the image patches, we obtained a dataset for each patient, with each row representing information about a single cell nucleus detected. However, our objective was to generate an integrated dataset that combines clinicoapathological data with image-related features. Therefore, we needed to aggregate the datasets to derive representative values for each patient. Hence, our approach was to cluster the cells nuclei together in groups with similar characteristics and to count for each patient the proportion of cells belonging to each different cluster. The clustering algorihtm chosen was FlowSOM, a two-level clustering technique based on self-organising maps, as it showed promising results in the analysis of cells markers [49]. The algorithm is presented in detail in Section 3.5.

Even if SOM can already be used to get a clustering, it gets advantageous to include more nodes in the grid than the expected number of clusters. Furthermore, the expected number of clusters of urothelium cells nuclei is unknown before-hand since this technique has never been used before. To overcome this problem, the node centers of the clusters are clustered together in this step to create meta-clusters. This second clustering approach is made by hierarchical clustering, as it provides detailed information about which observations are most similar to each other. This approach operates through multiple subsampling iterations of the points, performing hierarchical clustering for each subsample. The final clustering is determined by assessing how frequently the same points are grouped together or not across these iterations. The optimal number of meta-clusters is derived by the analysis of the dendrogram, an unbiased diagram that shows the hierarchical relationship between objects. By performing these steps, we were able to derive a final optimal number of meta-clusters that subdivided our nuclei population in smaller groups.

4.4. AI models

An artificial neural network called neuro-fuzzy modelling has been developed as a result of our systematic review study. The algorithm is explained in detail in Section 3.2. In addition, five machine learning models have been implemented, namely random forest, gradient boosting, support vector machine, decision tree and extreme trees. These traditional techniques fall under the category of supervised learning, where the data at hand comprises labeled instances, implying that each data point includes both features and a corresponding label. The algorithms are presented in Section 3.3.

4.4.1. Neuro-fuzzy modelling

We considered Gaussian membership functions, as they reported the highest performances in [41]. The function has the standard Gaussian form:

$$N(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where σ^2 is the variance and μ is the mean. Both σ and μ are trainable parameters, randomly initialised.

To achieve the optimal final classification, a binary cross entropy loss function has been selected for the neural network. This particular choice was made due to its ability to effectively handle binary classification problems, ensuring accurate predictions. To ensure efficient optimization of the loss function, an Adam optimizer with a learning parameter of $\alpha = 0.01$ was employed. The Adam optimizer is renowned for its capability to adaptively tune the learning rate during the training process, facilitating effective convergence to an optimal solution. Moreover, to prevent overfitting, an L1 regularization parameter was thoughtfully included in the objective function. This regularization technique aids in constraining the neural network's weights during the training process, enhancing the generalization in the test set. With the L1 regularization parameter embedded in the objective function, the network is able to learn representations from the data while minimizing the impact of noise or outliers. Additionally, to assess the model's performance and ensure its robustness, a 10-fold cross validation technique was performed. This involved dividing the available dataset into ten different subsets or folds, with the training and testing operations repeatedly conducted for each fold. This approach aids in obtaining a more reliable

estimation of the model's performance, reducing the likelihood of any biased assessment or accidental inconsistencies.

4.4.2. Machine learning models

In our study, we conducted an extensive hyperparameter tuning search to optimize the performance of our models. This process involved systematically exploring various parameters related to the construction of the models. One of the crucial parameters we focused on was the number of estimators, which refers to the number of trees in the ensemble. By fine-tuning this parameter, we aimed to find the optimal balance between underfitting and overfitting. Increasing the number of estimators may enhance the model's performance; however, too many estimators can lead to an overcomplicated model that fails to generalize well to unseen data. Additionally, we optimized the maximum depth parameter, which determines the maximum number of levels that a decision tree can possess. A deeper tree can capture more intricate patterns in the data, but it may also learn from the noise in the data. Moreover, we explored the maximum number of features per split, which defines the number of features considered when searching for the best split at each internal node. By experimenting with this parameter, we aimed to control the tree's diversity and avoid potential biases caused by specific features dominating the splitting process.

Through this extensive hyperparameter tuning search, we aimed to optimize the performance of our models by carefully selecting values for parameters such as the number of estimators, maximum depth, minimum sample split, and maximum number of features per split. By making informed choices about these parameters, we aimed to achieve models that accurately capture complex relationships in the data while avoiding overfitting and improving generalization capabilities.

To conduct the variable importance analysis, we have employed a permutation importance algorithm. In order to ensure the reliability and accuracy of our findings, we have opted for a substantial number of permutations. Specifically, we have chosen a large value of $K = 1000$ permutations for each variable. This high number of permutations allows for a comprehensive exploration of the possible variable combinations and their subsequent impacts on the analysis. Hence, it is possible to derive more accurate and trustworthy variable importance analysis results.

4.4.3. Implementation

This study was implemented using the Python programming language, leveraging key libraries such as TensorFlow and Scikit-learn. In addition, the NumPy and Pandas libraries were utilized to handle numerical computations and data manipulation, respectively. The code developed for this study is available for further analysis within Erasmus MC, facilitating future research.

5 Results

5.1. Dataset

The final analysis comprised a total of 900 patients after applying four exclusion criteria. For a visual representation of the exclusion criteria and patient selection process, refer to Figure 5.1, which illustrates the CONSORT chart. This chart provides a clear overview of the steps taken to arrive at the final cohort.

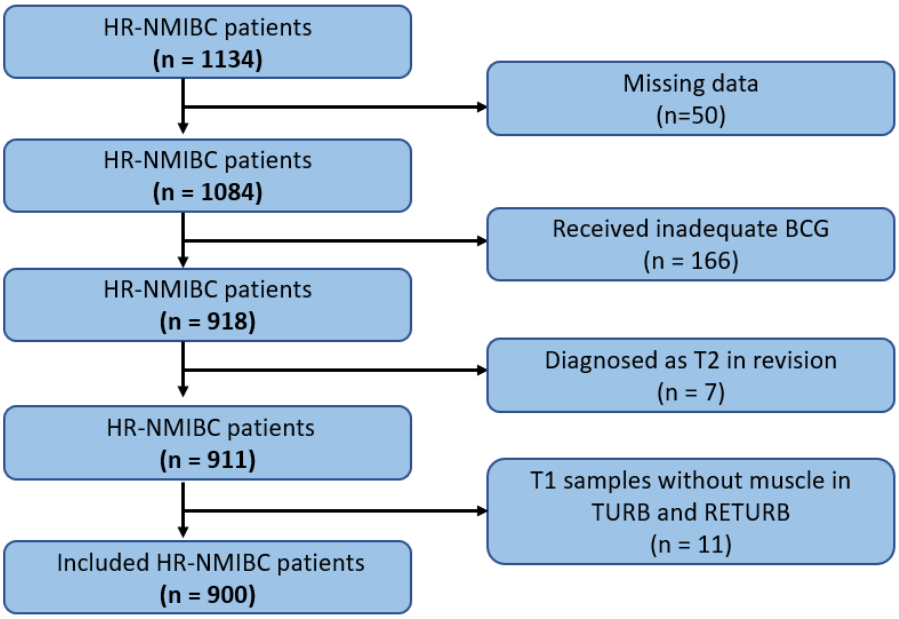

Figure 5.1: CONSORT chart of clinicopathological data cohort

A summary overview of the dataset, presented in Table 5.1, provides insights into the patient demographics and characteristics. The dataset was randomly partitioned into training and test sets using an 80-20% split ratio. The cohort was primarily composed of male individuals, with advanced age and a high prevalence of smokers. It is worth noting that the median follow-up period for the patients included in the dataset was 60 months. This duration of observation allows for a comprehensive evaluation of patient outcomes, including disease progression, recurrence, and other relevant clinical endpoints. Moreover, Table A.1 presents an overview of the training set. It can be noticed that the percentages remain consistent even after the split into train and test subsets. This finding ensures an unbiased selection process.

Clinical Parameter	Subgroup	All	BCG treatment		Progression		HG recurrence	
			Responders	Failure	Yes	No	Yes	No
ALL		900	659 (73%)	241 (27%)	132 (15%)	768 (85%)	144 (16%)	756 (84%)
Gender	Male	723 (80%)	523 (72%)	200 (28%)	110 (15%)	613 (85%)	119 (16%)	604 (84%)
	Female	177 (20%)	136 (77%)	41 (23%)	22 (12%)	155 (88%)	25 (14%)	152 (86%)
Age (years)	Median (min-max)	72 (32-100)	72 (32-98)	73 (45-100)	72.5 (41-92)	72 (32-100)	73 (45-92)	72 (32-100)
Smoking	Yes	552 (61%)	401 (73%)	151 (27%)	75 (14%)	477 (86%)	84 (15%)	468 (85%)
	No	348 (39%)	258 (74%)	90 (26%)	57 (16%)	291 (84%)	60 (17%)	288 (83%)
Size (cm)	≤ 3	832 (92%)	607 (73%)	225 (27%)	122 (15%)	710 (85%)	134 (16%)	698 (84%)
	> 3	68 (8%)	52 (76%)	16 (24%)	10 (15%)	58 (85%)	10 (15%)	58 (85%)
Staging	Tis	96 (11%)	70 (73%)	26 (27%)	12 (13%)	84 (87%)	16 (17%)	80 (83%)
	Ta	317 (35%)	255 (80%)	52 (20%)	45 (14%)	272 (86%)	36 (11%)	281 (89%)
	T1	487 (54%)	334 (69%)	153 (31%)	75 (15%)	412 (85%)	92 (19%)	395 (81%)
Grading	G1L	15 (2%)	13 (87%)	2 (13%)	0 (0%)	15 (100%)	1 (7%)	14 (93%)
	G2L	40 (4%)	32 (80%)	8 (20%)	9 (23%)	31 (77%)	3 (8%)	37 (92%)
	G2H	58 (6%)	46 (79%)	12 (21%)	13 (22%)	45 (78%)	5 (9%)	53 (91%)
	G3H	787 (87%)	568 (72%)	219 (18%)	110 (14%)	677 (86%)	135 (17%)	652 (83%)
Concomitant CIS	Yes	47 (5%)	31 (66%)	16 (34%)	21 (45%)	26 (55%)	7 (15%)	40 (85%)
	No	853 (95%)	628 (74%)	225 (26%)	111 (13%)	742 (87%)	137 (16%)	716 (84%)
History of cancer	Yes	202 (22%)	157 (78%)	45 (22%)	27 (13%)	175 (87%)	29 (14%)	173 (86%)
	No	698 (77%)	502 (72%)	196 (28%)	105 (15%)	593 (85%)	115 (16%)	583 (84%)
Number of tumors	Single	382 (42%)	294 (75%)	88 (25%)	35 (17%)	347 (83%)	52 (14%)	330 (86%)
	Multiple	280 (31%)	194 (69%)	86 (31%)	29 (10%)	251 (90%)	58 (21%)	222 (79%)
Follow-up (months)	Median (min-max)	60 (2-228)	72 (2-228)	60 (7-204)	48 (7-180)	72 (2-228)	48 (3-180)	72 (2-228)

Table 5.1: Dataset overview

Whole-slide images obtained from patients who were excluded based on clinical considerations were omitted from the analysis. Finally, we also excluded images that exhibited punching artifacts. The CONSORT chart of the exclusion criteria is presented in Figure 5.2.

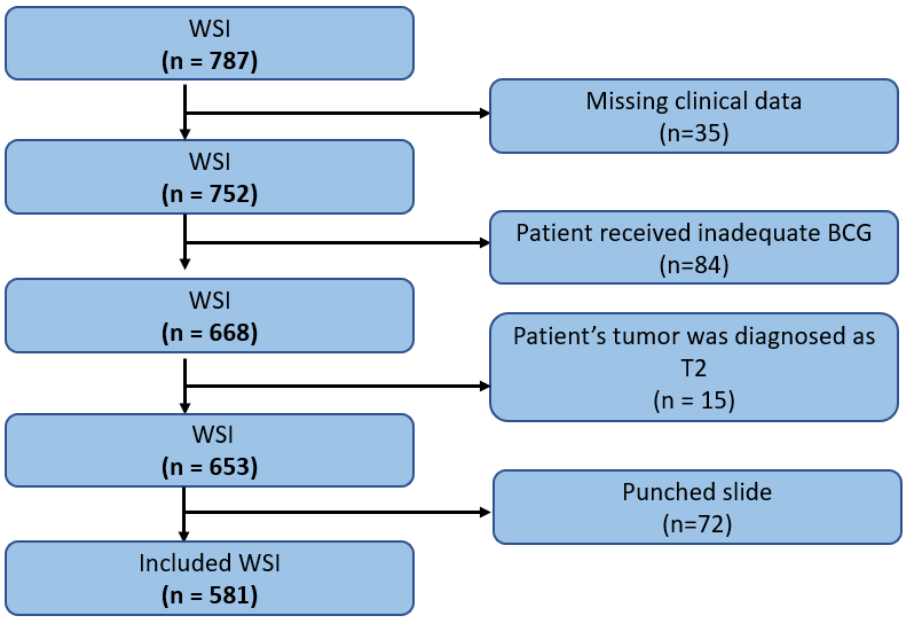

Figure 5.2: CONSORT chart of WSI cohort

As a result, a total number of 581 whole slide images have been included. For each patient, one or multiple slides were available. The final number of slides included correspond to 504 HR-NMIBC patients.

5.2. Image segmentation

In order to analyze the gigapixel whole slide images, we employed a methodology inspired by the technique described by [59]. This involved subdividing the large images into smaller patches that specifically contained urothelium tissue. The resulting output consisted of thousands of small patches,

each measuring 512 x 512 pixels, which were subsequently available for segmentation. To establish a reference for evaluating the segmentation model, we created a ground truth dataset consisting of 200 images. This ground truth served as a benchmark against which we could assess the performance of the segmentation model. An example of ground truth can be found in Figure 5.3.

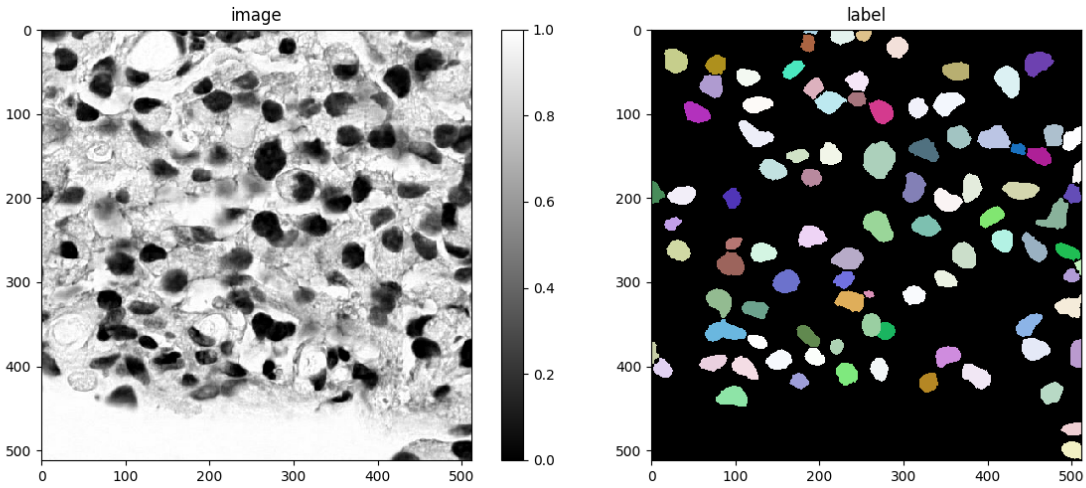

Figure 5.3: Example of an image patch (left: image patch, right: ground truth created)

For this purpose, we utilized the pretrained model named *2D_versatile_he*, developed by [49]. However, the initial results obtained from applying the pretrained model, presented in Figure B.1, were not satisfactory for accurate segmentation. The model exhibited an accuracy level of approximately 60%. As a result, we made the decision to undertake retraining of the model using our own images and corresponding ground truth data. This approach aimed to enhance the model's performance. The outcomes of this retraining process are presented in Figure 5.4, which showcases the improved performance achieved by the CNN. By retraining the model using our own dataset and ground truth labels, we were able to enhance the accuracy and segmentation performance.

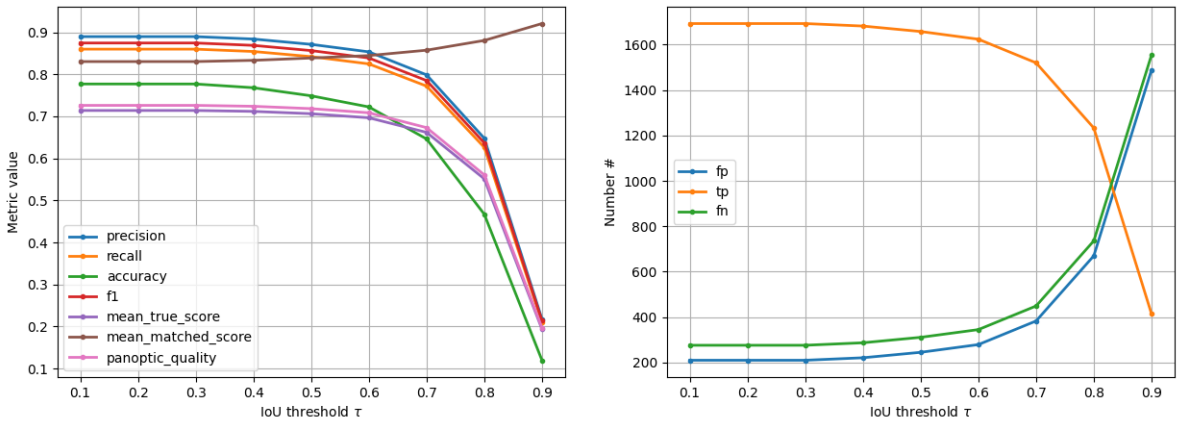

Figure 5.4: Performances of nuclei segmentation with Stardist

The evaluation of segmentation performance is conducted based on various metrics, considering the Intersection over Union (IoU) threshold τ. Through careful analysis, we determined that an optimal value of $\tau = 0.46$ yielded the optimal segmentation results. At this threshold, the segmentation exhibited high performance metrics, along with a low number of false positives and false negatives. The F1 score, a measure of the model's accuracy, reached 87%, indicating the overall balance between precision and recall. The precision was measured at 88%, while the recall stood at 85%. Therefore, we saved the

weights and biases of the model. This enabled us to apply the trained algorithm to segment the entire cohort of images, providing comprehensive and accurate segmentation across the dataset. As a visual representation of the segmentation output, Figure 5.5 showcases an example of an image that has been segmented using the retrained model.

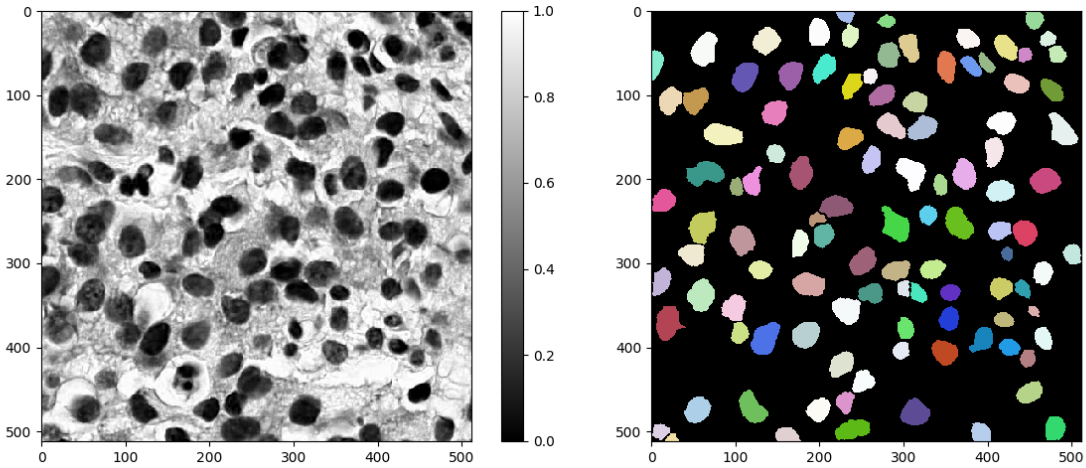

Figure 5.5: Example of segmentation with Stardist (left: original image, right: segmented image)

The employed technique successfully facilitated the detection of the cells nuclei within the patches. To avoid the detection of lymphocites and other immune cells, we set a minimum treshold for the area of $25\mu m^2$, as around 97.5% of the nuclei were found to be smaller than the treshold [61]. Following the segmentation process, we computed a total of 22 features for each detected object. These features encompassed a range of key properties, including area, perimeter, eccentricity, solidity, and various others. Each of these features provided valuable insights into the morphology and characteristics of the nuclei. To organize and capture the information pertaining to the nuclei morphology, we created a dataframe for each image and each patient. This dataframe contained the computed features as individual columns. The inclusion of multiple features provided a rich set of information, enabling a detailed characterization of the nuclei properties within each image and patient.

5.3. Clustering

The underlying concept of our approach involves grouping nuclei with similar characteristics, enabling us to analyze and interpret their collective behavior. To achieve this, we adopted the FlowSOM clustering technique, which facilitated the aggregation of the millions of cell nuclei features that had been extracted. Initially, we trained a self-organizing map (SOM) on the unlabeled data, resulting in the creation of a 10 x 10 grid of nodes. Each node within the SOM grid represented a distinct cluster, grouping together nuclei that exhibited similar properties or feature patterns. This process allowed us to capture and define the heterogeneity present within the dataset. Subsequently, we constructed a minimum spanning tree (MST) on top of the SOM grid. The MST is a graph that connects the nodes of the SOM grid, highlighting the relationships and similarities between different clusters. The resulting MST graph provides a visual representation of the interconnections and structure of the clusters, enabling us to better understand their associations and dependencies. Figure 5.6 illustrates the resulting MST graph. The visualization displays each node of the original SOM grid. To provide further information, an additional colored circle surrounds each node, representing the meta-cluster to which it belongs. The angles represent the mean cluster values of each node in star charts.

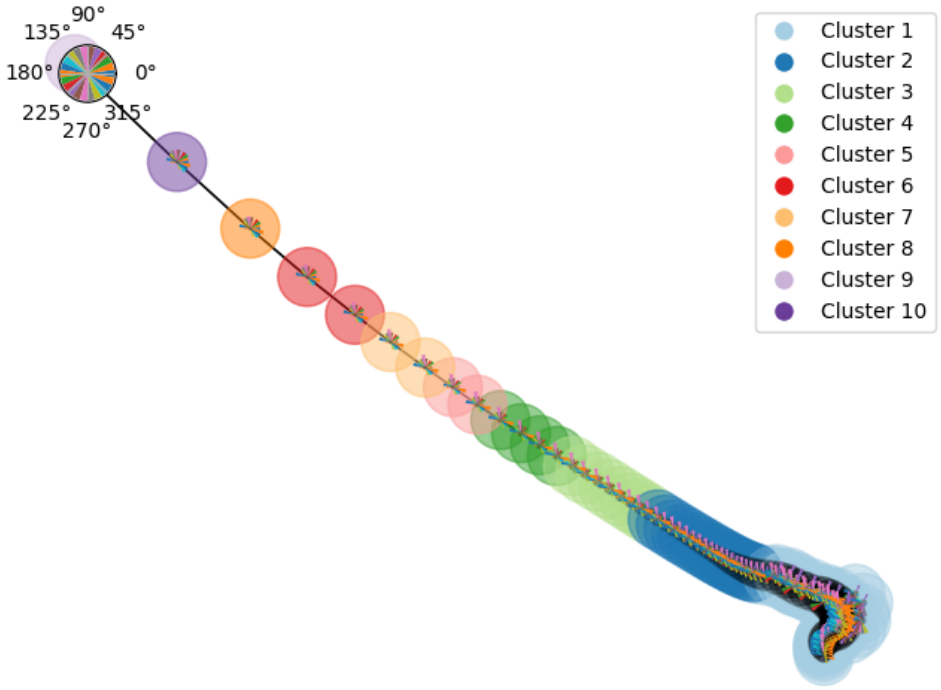

Figure 5.6: Minimum spanning tree

The objective of this particular step is to establish connections between nodes in the grid that exhibit similarity in order to construct a tree-like structure. This structure is designed to minimize the distances between the connected nodes, enhancing the representation of their relationships. Upon examination of the resulting tree structure, it becomes apparent that the majority of nodes are concentrated in the right portion of the tree. As one traverses towards the left on the tree, the number of nodes progressively decreases. This observed trend signifies a potential clustering pattern or organization within the data, with more closely related nodes clustered on the right side of the tree. The clustering of nodes in the right region of the tree suggests the presence of groups that share similar characteristics or properties.

The last step of our technique involves the creation of meta-clusters based on the structure obtained from the previous steps. Meta-clustering involves generating clusters from a diverse set of individual clusterings, integrating the information obtained from various sources. To achieve this, we employ a hierarchical clustering algorithm, which enables the formation of cohesive groups based on the relationships identified within the data. Determining the appropriate number of clusters for the meta-clustering step poses a challenge, as the expected number is unknown a priori. To address this, we implemented an unbiased technique known as the dendrogram. The dendrogram is a graphical representation that depicts the hierarchical relationships between objects, with the heights of the branches reflecting the distances between the clusters. The dendrogram serves as a valuable tool for selecting the optimal number of clusters in a data-driven manner. By visually analyzing the dendrogram, we can identify distinct branches or levels where the distances between clusters exhibit significant changes. These points of variation indicate potential divisions into separate meta-clusters. Figure 5.7 presents the dendrogram diagram, providing a visual representation of the hierarchical relationship between the clusters.

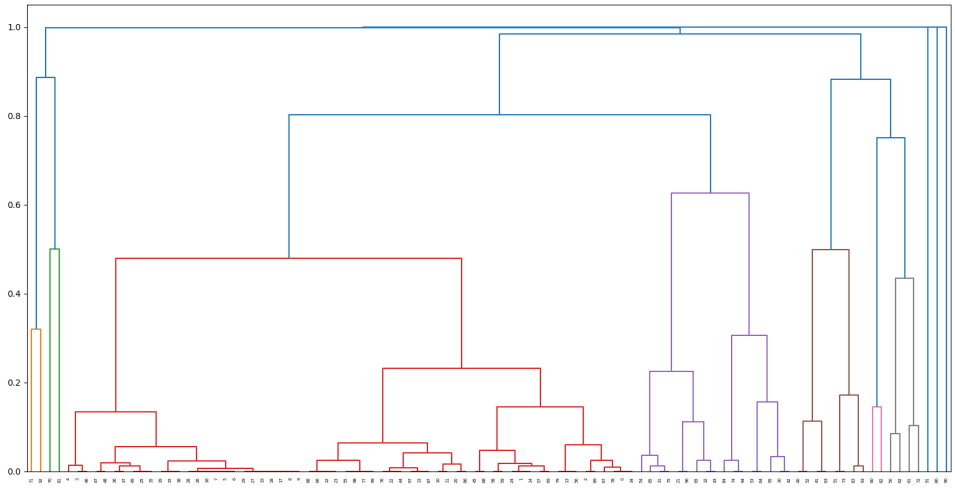

Figure 5.7: Dendrogram

Upon examination of the dendrogram, a crucial step in determining the optimal number of meta-clusters is to identify the highest vertical gap that remains unobstructed by any horizontal line. In this particular dendrogram, we observe such a vertical gap that ranges between y=0.63 and y=0.73. By counting the number of vertical lines that cross this gap, we can deduce the optimal number of meta-clusters to be ten. Specifically, we find that ten vertical lines intersect this vertical gap, indicating that this number of clusters provides the most meaningful and distinct divisions within the dataset. This approach, based on the identification of the highest unobstructed vertical gap, ensures an unbiased determination of the optimal number of meta-clusters. By leveraging this method, we can establish the appropriate number of clusters to accurately capture the underlying structure and heterogeneity within the dataset.

In Figure 5.8, we showcase the final distribution of cells among the ten distinct clusters obtained through our analysis. To facilitate visualization, the y-axis is presented in a logarithmic scale. Each bar within the plot represents the percentage of the total number of cells attributed to each respective cluster. A noteworthy observation is that Cluster 1 encompasses the largest majority of cells, indicating a prominent representation within the dataset. The second cluster comprises approximately 9% of the total number of nuclei, signifying a significant but smaller proportion compared to Cluster 1. All the remaining clusters, contain less than 1% of the total number of cells, highlighting their relatively smaller presence within the dataset.

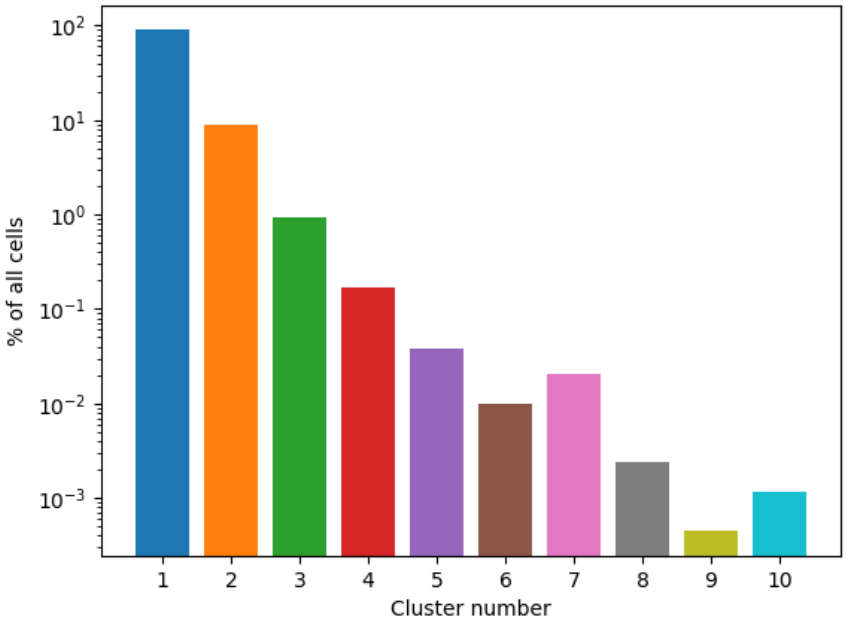

Figure 5.8: Cells distribution among the clusters

Following the clustering analysis, we proceeded to calculate the percentage of cells assigned to each cluster for every individual patient. This computation provided us with a quantitative measure of the distribution of cells among the different clusters for each patient. These percentages, representing the relative contribution of each cluster, served as new features for subsequent analysis. By incorporating these features alongside the clinicopathological data, we aimed to create an integrated dataset that encompassed both cellular information and relevant clinical variables. This integrated dataset, combining the clinicopathological data with the cluster centroid features, allowed us to perform a comprehensive analysis that considered both the cellular composition and the clinical context.

5.3.1. Clusters analysis

In our study, we employed a clustering technique to partition the cohort of cell nuclei into ten distinct clusters, aiming to achieve an optimal subdivision. This clustering analysis allowed us to uncover underlying patterns and groupings within the dataset. Figure 5.9 visually presents a normalized heatmap, showcasing the most relevant nuclei features for each cluster. In the heatmap, each row corresponds to a different feature, while the clusters are represented as columns. The intensity of the color reflects the magnitude of the feature value, with brighter colors indicating higher values. Upon closer examination, it becomes evident that clusters 8, 9, and 10 consistently exhibit the highest values across various features. However, it is noteworthy that the eccentricity of the nuclei exhibits a reversed pattern. Surprisingly, cluster 2 outperforms other clusters in terms of eccentricity, ranking first among them. This finding highlights the distinctive nature of cluster 2 in terms of nuclei eccentricity, suggesting its potential significance in capturing specific cellular characteristics.

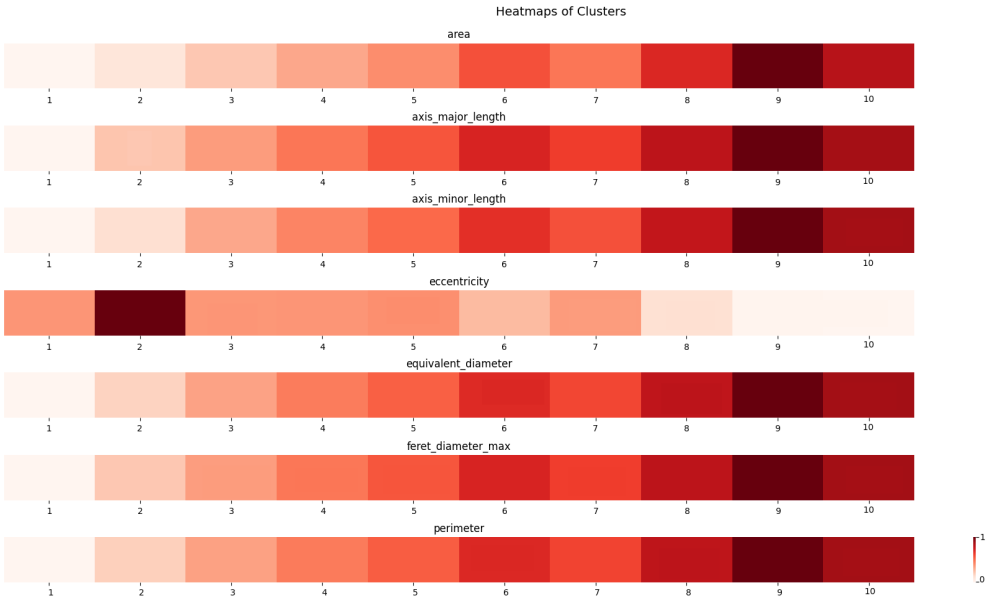

Figure 5.9: Heatmap of all the clusters

Figure 5.10 presents a boxplot depicting the distribution of nuclei area across the various clusters. Of particular interest is cluster 9, which has shown great promise in our analysis. Notably, the boxplot reveals that the nuclei within cluster 9 exhibit larger areas compared to almost all other clusters. This observation suggests that cluster 9 contains a subgroup of cells characterized by larger nuclear sizes. Examining the overall distribution, we observe that the majority of cells, predominantly grouped within cluster 1, tend to have smaller nuclei areas compared to other clusters. In other words, the analysis indicates that a significant portion of the cells present in the dataset are relatively small in size. Consequently, the area of the nuclei emerges as one of the most distinguishing factors contributing to the clustering process.

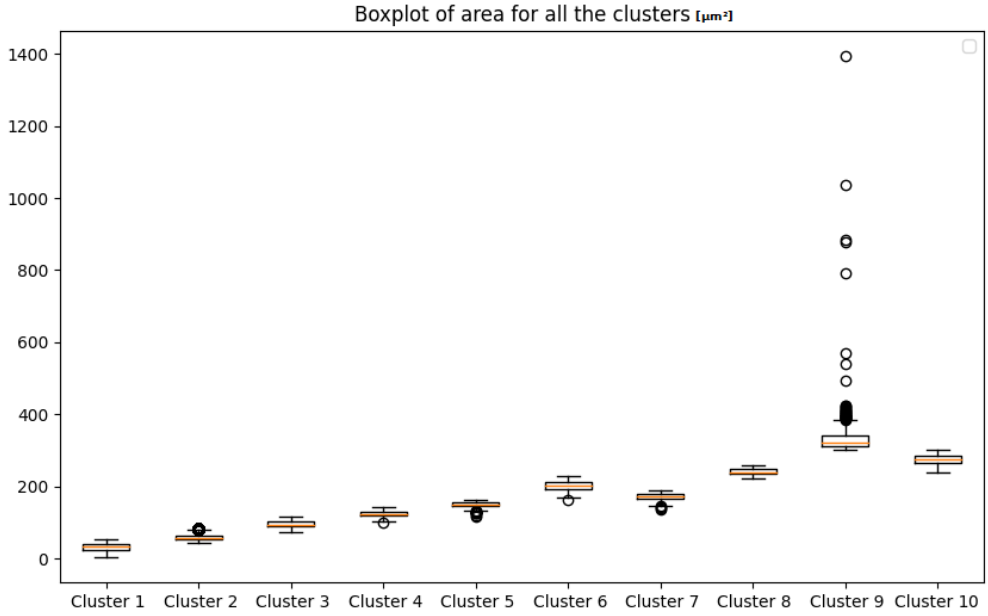

Figure 5.10: Boxplot of area

An additional crucial feature to analyze is the eccentricity of the cells. Eccentricity measures the extent to which a curve deviates from the circularity of the given shape. Values closer to zero indicate a more circular object, while higher values signify a greater elongation of the object. Figure 5.11 displays the distribution of this feature across the clusters in our cohort. Upon examination, it becomes apparent that cluster 1, which is also regarded as one of the most promising clusters in terms of predictive performance, exhibits higher eccentricity values within the interquartile distribution. This finding suggests that the cells within this cluster possess a greater elongation compared to other cell types.

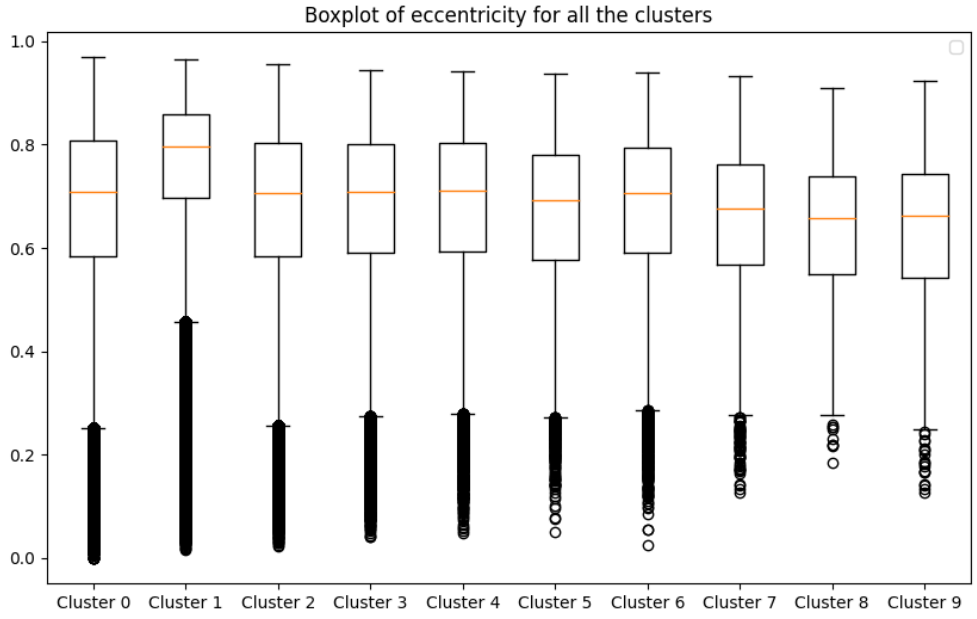

Figure 5.11: Boxplot of eccentricity

As a final step in our analysis, we sought to investigate whether there were any statistical differences in the distributions of cells among clusters for different patients. Specifically, we divided the patient cohort based on different endpoints (e.g., progressors versus non-progressors) and examined whether the nuclei features exhibited any variations between these two groups. Surprisingly, this analysis did not reveal any significant differences, suggesting that the types of cells within each cluster are generally similar across different patients. To further explore potential distinctions among patients, we turned our attention to the abundance of cell types within each group. We once again divided the patients based on the endpoints and calculated the distribution of cell type abundances. Figure 5.12 provides an example of the abundance distribution for cluster 10. The y-axis indicates the amount of cells. We can observe that progressors tend to exhibit a slightly higher percentage of cells belonging to cluster 10 compared to non-progressors. However, it is important to note the presence of outliers and the similarity in the distribution of cell abundances. These factors limit the strength of these findings and indicate that further investigation is warranted to draw definitive conclusions.

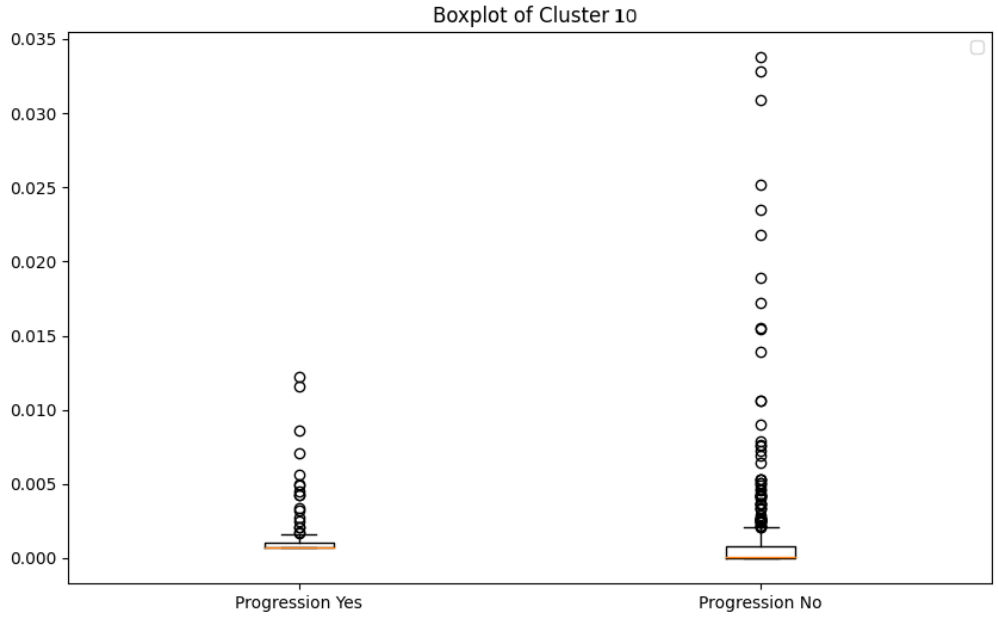

Figure 5.12: Boxplot of abundance for cluster 10

5.4. Machine learning models performance

To predict the binary endpoints of progression, BCG failure, and HG recurrence, we employed six different classifiers. These classifiers were designed to analyze the dataset and make predictions based on distinct features and algorithms. In order to assess the performance of these models, we utilized five commonly used statistical metrics that evaluate the models' ability to accurately classify patients. To establish a baseline for comparison, we initially conducted an analysis using only clinicopathological data from a subset of 504 patients. This baseline analysis served as a reference point to gauge the effectiveness of subsequent models. Our goal was to improve upon these baseline results using alternative approaches: utilizing image features alone, integrating image features with clinicopathological data, and expanding the analysis to include the entire cohort of 900 patients. Comprehensive tables displaying the performance metrics for all the models can be found in Tables B.1 to B.12. These tables provide a detailed overview of the models' performance, enabling a comprehensive comparison of their respective abilities to accurately classify patients. Among the different classifiers employed, the random forest model consistently demonstrated superior performance, exhibiting high discriminative ability across all the evaluated metrics. This model showcased promising results in accurately predicting the binary endpoints of progression, BCG failure, and HG recurrence. The robustness and effectiveness of the random forest model make it a valuable tool for classification tasks within this dataset.

The prediction of BCG failure yields a baseline performance of approximately 70% across the considered metrics, as demonstrated in Table 5.2. Similarly, employing only image-related features produces comparable results. However, notable enhancements can be observed when utilizing the integrated dataset comprising clinicopathological data and image-related features as input. The analysis incorporating clinicopathological data for the entire cohort leads to a slight decrease in the metrics.

Table 5.2: BCG failure

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
RF	Clinical data (n=504)	0.70	0.72	0.70	0.77	0.63
RF	Image features (n=504)	0.69	0.71	0.69	0.75	0.65
RF	Clinical data and image features (n=504)	0.77	0.77	0.77	0.80	0.73
RF	Clinical data (n=900)	0.66	0.68	0.66	0.71	0.61

Table 5.3 showcases the performance of the random forest model in predicting disease progression. The initial baseline analysis exhibits similar performance to that of BCG failure. Nevertheless, it becomes evident that the inclusion of the integrated dataset significantly enhances the methods' discriminative power, resulting in metrics reaching scores of approximately 80%. Moreover, expanding the study to include more patients improves model performance and yields discriminative power comparable to that of the analysis conducted using the integrated dataset.

Table 5.3: Progression

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
RF	Clinical data (n=504)	0.68	0.69	0.70	0.73	0.67
RF	Image features (n=504)	0.70	0.71	0.70	0.74	0.67
RF	Clinical data and image features (n=504)	0.78	0.80	0.78	0.84	0.72
RF	Clinical data (n=900)	0.77	0.76	0.77	0.75	0.74

Lastly, Table 5.4 presents the metrics for the classifier in predicting high-grade recurrence. In this instance, the models generally perform worse than the previous endpoints, with baseline values of approximately 60%. However, the integration of image-related features with clinicopathological data evidently improves the results. The integrated analysis demonstrates metrics of around 70% across all methods. Furthermore, the inclusion of new patients slightly enhances the baseline models.

Table 5.4: HG recurrence

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
RF	Clinical data (n=504)	0.61	0.62	0.61	0.63	0.58
RF	Image features (n=504)	0.60	0.60	0.60	0.62	0.59
RF	Clinical data and image features (n=504)	0.71	0.72	0.71	0.74	0.68
RF	Clinical data (n=900)	0.63	0.66	0.63	0.71	0.54

5.5. Variable importance

After conducting an analysis of the classifiers, we employed a permutation importance method to gain insights into the significance of various features in the prediction process. This technique involved randomly shuffling the values within each column of the dataset a total of 1000 times, enabling us to ascertain the relative contribution of each feature. By systematically permuting the values and observing the resulting impact on the predictive performance, we were able to identify the features that played the most substantial role in the prediction task. This approach allowed us to assess the importance of each feature by measuring the changes in performance metrics following the shuffling process.

In the context of predicting BCG failure, Figure 5.14 provides a visual representation of the process involved. The figure shows that the first cluster holds the highest level of relevance, followed closely by cluster 10. Within these clusters, certain clinicopathological factors emerge as the most influential in determining the outcome. Upon closer inspection, it becomes apparent that smoking and age exert the most significant impact among the clinicopathological factors considered. These variables hold substantial importance in the prediction of BCG failure, potentially serving as key indicators or risk factors associated with the outcome.

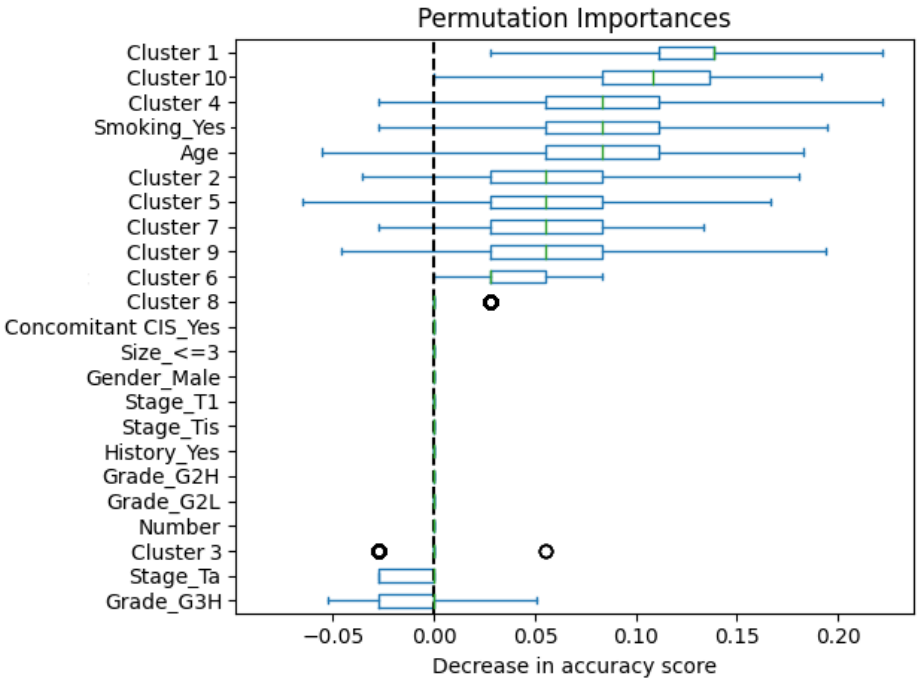

Figure 5.13: Variable importance progression

The permutation importance for predicting progression is illustrated in Figure 5.13. An analysis of the figure reveals that among the various clusters, cluster 4 emerges as the most influential in driving the predictive outcome, showcasing its importance in this context. Following closely behind cluster 4, the age of the patient stands out as another crucial factor that significantly impacts the prediction of progression.

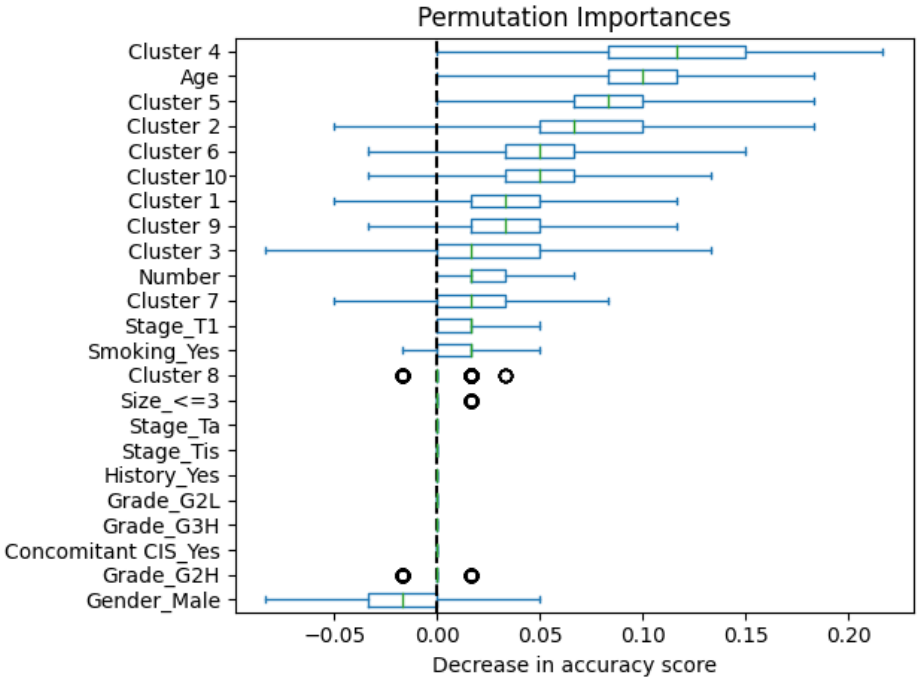

Figure 5.14: Variable importance BCG failure

The results for the high-grade recurrence endpoint are showcased in Figure 5.15. Notably, cluster 10 emerges as the most decisive cluster in determining the likelihood of high-grade recurrence. Following cluster 10, group number four exhibits noteworthy significance in the prediction process. This implies that the features encompassed within this group contribute significantly to the accuracy of the high-grade recurrence prediction model.

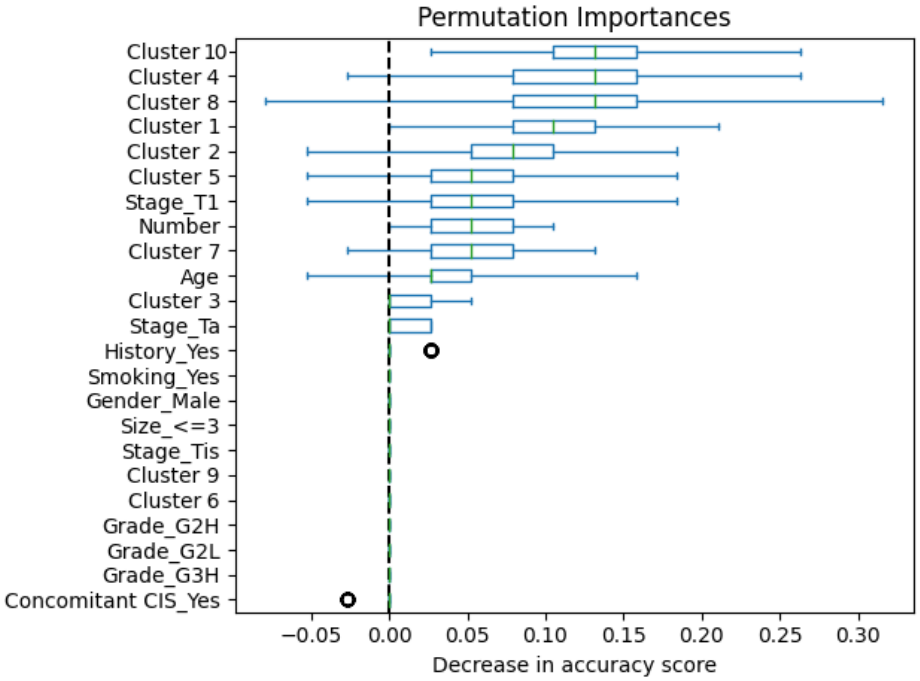

Figure 5.15: Variable importance HG recurrence

6

Discussion and conclusions

Through our research, we have discovered that integrating clincopathological data with image-related features enhances the performances of artificial intelligence techniques in predicting the clinical outcome of HR-NMIBC. Our work has resulted in the development of a novel approach that effectively combines the information from cells nuclei per patient, allowing for improved classification accuracy. By combining clincopathological data with image-related features extracted from medical imaging, we have been able to show the power of AI to more accurately predict the clinical outcome of HR-NMIBC. This integrated analysis approach provides a comprehensive view of the disease, taking into account both the macroscopic characteristics and the microscopic details observed in the cellular level. Our research has led to the successful proposal of a methodology that aggregates the information obtained from thousands of cell nuclei for each patient. This aggregation process transforms the raw data into a format that can be effectively used by a classifier, a machine learning algorithm designed to categorize and predict outcomes. By using this combined information, we improved the performance of AI techniques in predicting the clinical outcome of HR-NMIBC.

The dataset overview (Table 5.1) reveals several imbalanced variables and potential biases that require consideration during the analysis and interpretation of the findings. Firstly, there is a gender imbalance, with a higher representation of males (80%) compared to females. This gender bias might impact the generalizability of conclusions, as the dataset may not adequately represent the characteristics and outcomes of both genders. The dataset also displays a difference in tumor size, with the majority of cases (92%) having a size of 3 cm or less, potentially affecting the generalizability of results for larger tumors. The majority of cases (87%) belong to the G3H grade, while G1L (2%), G2L (4%), and G2H (6%) have lower representation. This finding poses a potential disproportion in grading levels that could impact the analysis of grade-outcome relationships. Furthermore, also the endpoints showed to be imbalanced. Around 73% of the cohort was in the group of BCG responders, whereas about 85% of the patients did not show any recurrence or progression.

We achieved progress by retraining a convolutional neural network using our own image dataset. This CNN model played a crucial role in our analysis and demonstrated remarkable performance in segmenting hematoxylin and eosin stained images. Our findings particularly highlighted the effectiveness of a model based on star-convex polygons in accurately segmenting cell nuclei, specifically in the context of urothelium cells.By retraining the CNN, we used its deep learning capabilities to extract valuable insights from the H&E stained images. The model exhibited high accuracy in identifying and delineating the boundaries of cell nuclei. This breakthrough in segmentation lays the foundation for subsequent analyses and investigations in our study. The success of our retrained CNN model in segmenting H&E stained images underscores its potential as a valuable tool in the field of histopathology and medical image analysis.

In addition to our achievements in retraining a convolutional neural network and utilizing a star-convex polygon model for cell nucleus segmentation, we also implemented an innovative clustering technique called FlowSOM. This technique proved to be highly effective in grouping millions of cell nuclei based

on their characteristics, and it demonstrated suitability for our specific research purposes. FlowSOM enabled us to harness the power of clustering to organize and classify the vast number of cell nuclei present in our dataset. By grouping together nuclei, we gained valuable insights into the heterogeneity and diversity within the cellular population. This technique allowed us to leverage the information from multiple nuclei within each patient, providing a more comprehensive representation of the underlying biology. What sets our approach apart is the novel method we devised for quantifying the nuclei belonging to each cluster for every patient. This counting approach, which had not been previously explored, proved to be an effective strategy for our specific task. We obtained valuable information about the cellular composition and heterogeneity within individual samples, by quantifying the distribution of nuclei across different clusters for each patient. The integration of FlowSOM into our analysis pipeline represented a crucial step forward, empowering us to handle the complex and immense amount of cellular data present in our study. By combining the segmentation accuracy of our retrained CNN model with the clustering capabilities of FlowSOM, we were able to extract meaningful information about the composition, organization, and characteristics of cell nuclei across a large cohort of patients.

During the final stage of our work, we implemented state-of-the-art AI models that exhibited relatively high performance metrics. Our findings align with previous studies discussed in the systematic review, further reinforcing their validity. Specifically, our technique achieved an accuracy rate of approximately 80% when assessed using various evaluation metrics. This robust performance serves as a strong indicator of the effectiveness and reliability of our approach, corroborating the results presented in the systematic review. The models, trained on our integrated dataset comprising clincopathological data, and information extracted from clustered nuclei, provided valuable insights and demonstrated notable predictive capabilities. Among the different algorithms tested, the random forest algorithm emerged as the most suitable for the task at hand. Random forest's ability to handle high-dimensional datasets, capture non-linear relationships, and reduce overfitting contributed to its superiority in our study.The success of our models not only highlights the effectiveness of our approach but also encourages further exploration and refinement of AI methodologies for other medical domains.

Furthermore, we conducted a comprehensive variable importance analysis to gain insights into the features that strongly influence the prediction outcomes. This analytical approach holds remarkable significance in the medical field, as merely achieving high accuracy without understanding the underlying factors driving the model's conclusions would render the results less meaningful. The variable importance analysis provided valuable insights into the specific characteristics of cell clusters that have the greatest impact on the prediction of clinical outcomes in HR-NMIBC. It can be noticed that Cluster 1, 4, and 10 stand out as the most important clusters, ranking first in predicting progression, BCG failure, and HG recurrence, respectively. Additionally, Cluster 10 achieves the second position in the progression analysis, while age emerges as the second most influential variable for BCG failure, and Cluster 4 takes the second spot for HG recurrence prediction. Notably, our findings revealed that clusters containing larger and more elongated cells exhibit a higher degree of relevance in determining the clinical outcome. This observation suggests that cellular size and shape play a significant role in disease progression and response to treatment. Our findings emphasize the significance of cellular morphology and highlight the potential relevance of cell size and shape as prognostic factors in HR-NMIBC. These insights have the potential to guide further research and investigations, enabling a deeper understanding of the disease mechanisms and facilitating the development of novel diagnostic and therapeutic approaches.

While our study has made significant strides in advancing the prediction of clinical outcomes in high-risk non-muscle invasive bladder cancer, it is crucial to acknowledge the limitations inherent in our research. These limitations warrant careful consideration and highlight areas for further investigation and improvement. Firstly, it is important to recognize that the study cohort used for this analysis consisted exclusively of HR-NMIBC patients. Therefore, the generalizability of our findings to other cohorts of NMIBC patients may be challenging. The specific characteristics and underlying biology of HR-NMIBC patients may differ from those with different risk profiles. Therefore, caution should be exercised when extrapolating our results to broader populations, emphasizing the need for additional studies encompassing diverse patient cohorts. Moreover, while the classifiers employed in our study demonstrated notable performance, it is essential to acknowledge that perfection has not yet been attained. While

our methods have shown suitability for the task at hand by providing valuable predictions and insights into determining factors, it is important to recognize that the classifiers' accuracy is not flawless. The trade-off between providing valuable insights and achieving higher accuracy is an important consideration. While our methods excel in providing interpretability and understanding of the predictive process, they may sacrifice a fraction of accuracy when compared to alternative methods solely focused on maximizing performance metrics.

In the pursuit of translating our methods into clinical applications, it is essential to delve deeper into the underlying biology of the distinct clusters identified in our study. Gaining a comprehensive understanding of the characteristics and functional implications of these clusters holds great potential in enhancing the clinical utility of our findings. While we have shown the primary characteristics of the most promising clusters, further investigation is needed to unravel their biological significance. Interestingly, the analysis of cells within subpopulations associated with different clinical endpoints did not reveal noticeable differences among the groups. This finding suggests that the distinguishing factors contributing to divergent clinical outcomes may not lie only within these subpopulations. The limited disparity observed in the cellular composition among the groups warrants further exploration to uncover the factors that play a more crucial role in discerning the differences between the clinical endpoints. Furthermore, exploring the differences in cell abundance among the groups adds an additional layer of complexity to our analysis. While these discrepancies have been observed, the precise implications and underlying reasons for such variations remain unclear. Understanding the factors that contribute to the variation in clinical outcomes can provide valuable insights into the disease mechanisms and potential therapeutic targets.

Our study has culminated in the development of a comprehensive pipeline for predicting the clinical outcome of high-risk non-muscle invasive bladder cancer. This pipeline encompasses multiple stages, starting with the creation of an integrated dataset that combines clinicopathological features and image analysis. This integrated dataset serves as the input for various classifiers, allowing us to evaluate their performance. To assess the effectiveness of our approach, we conducted comparative analyses using different datasets. This included datasets comprised solely of clinicopathological features or image-related features. Remarkably, our findings consistently demonstrated that the integrated analysis, incorporating both types of data, outperformed the individual feature-based analyses. This robust performance highlights the value of combining diverse information sources for accurate clinical outcome predictions in HR-NMIBC.

Moreover, through our investigation, we gained valuable insights into the inner workings of the predictive models. Specifically, we discovered that clusters of cells characterized by larger size and greater elongation appeared to have a more substantial influence on the predictions. This observation suggests that these clusters hold key insights into the underlying biology and potential drivers of clinical outcomes. However, further exploration is needed to explain the precise biological mechanisms at play within these clusters, as they may reveal previously unknown factors that impact the clinical outcome. Understanding the biological significance of these clusters can have profound implications for introducing our study into a clinical setting. By unraveling the unknown factors associated with specific cellular clusters, we can enhance the interpretability and clinical relevance of our predictions. This deeper understanding can guide clinicians in making informed decisions and aid in tailoring personalized treatment strategies for HR-NMIBC patients.

In conclusion, our study presents a robust pipeline for predicting the clinical outcome of HR-NMIBC, leveraging an integrated dataset and employing various classifiers. The integration of clinicopathological features and image analysis proved to be superior to individual feature-based analyses. Additionally, our investigation highlighted the importance of specific cellular clusters and their characteristics in driving predictive outcomes. Further exploration of the underlying biology of these clusters holds promise for uncovering novel factors and advancing the clinical applicability of our study in HR-NMIBC management.

7

Future research

In addition to the significant achievements made thus far, there are several avenues for further extending and enhancing this research. One possibility lies in expanding the scope of image analysis beyond the geometrical features of individual nuclei that were previously explored. While our focus primarily revolved around individual nuclei as distinct entities, there is potential in investigating the interactions and relationships between nuclei that are in close proximity to one another. By delving into the analysis of nuclear interactions, we can reveal hidden patterns and spatial distributions that might not be apparent through the analysis of isolated nuclei alone. Exploring the collective behavior and organization of nuclei within cellular clusters can provide valuable insights into the underlying biological processes. Analyzing the spatial relationships between nuclei may reveal important information about cellular interactions. These insights have the potential to deepen our understanding of the disease progression mechanisms and shed light on critical factors that impact clinical outcomes in HR-NMIBC. To pursue this avenue of research, advanced image analysis techniques, such as spatial statistics or graph theory, can be employed. These methodologies enable the quantitative assessment of spatial patterns, connectivity, and clustering of nuclei within tissue sections.

Expanding our research to incorporate the analysis of nuclear interactions alongside additional types of data holds potential for advancing our understanding. In particular, the inclusion of mitotic figures in the analysis can provide valuable insights into the proliferative activity and cellular dynamics within the tumor microenvironment. Mitotic figures, representing cells in the process of cell division, are important indicators of cellular proliferation and tumor aggressiveness. By incorporating the count of mitotic figures into our analysis, we can gain a more comprehensive understanding of the cellular activity and growth patterns within HR-NMIBC. Moreover, combining our data with spatial information and nuclear interactions can further enhance our understanding of HR-NMIBC at a molecular level. By integrating diverse data types, we can unravel complex relationships between genomic alterations, spatial heterogeneity, and nuclear interactions. Furthermore, it provides a foundation for the development of personalized treatment strategies that target specific molecular alterations and cellular interactions. To collect these data, advanced computational approaches are necessary. Deep learning models can be employed to recognise and count the mitotic figures in a whole slide image.

While our current approach successfully employed clustering techniques to group cells based on shared characteristics, it is worth considering other methodologies that can offer complementary perspectives. For instance, multiple instance learning (MIL) presents a compelling alternative. MIL focuses on the classification of sets, or bags, of instances rather than individual instances. In the context of HR-NMIBC, MIL could be applied to classify sets of cells, capturing the inherent heterogeneity and interplay among cells within a tissue sample. By taking into account the collective behavior of cells within a bag, MIL can reveal important patterns and dynamics that might not be evident when analyzing cells in isolation. Furthermore, attention-based methods represent another intriguing avenue to explore. Attention mechanisms allow models to selectively focus on certain regions of interest within whole slide images. By incorporating attention mechanisms into our analysis, we can identify the specific regions that have the greatest influence on the predictions. This provides valuable interpretability and can guide patholo-

gists and clinicians in understanding the salient features and regions that contribute to the HR-NMIBC prediction. These approaches can provide insights into the spatial distribution of cellular clusters, and highlight critical regions within whole slide images. By leveraging the power of these techniques, we can gain a more comprehensive understanding of the complex spatial dynamics and heterogeneity within HR-NMIBC tumors.

In summary, extending our research to include the analysis of nuclear interactions represents a promising avenue for further investigation. By exploring the spatial relationships and collective behavior of nuclei within cellular clusters, we can uncover hidden patterns, gain insights into cellular dynamics. The integration of nuclear interactions with additional data types, such as the count of mitotic figures, represents a powerful approach to enhance our understanding of HR-NMIBC. By incorporating these diverse datasets, we can decipher the intricate interplay between genetic alterations, spatial organization, and cellular dynamics, leading to a deeper comprehension of the disease's molecular mechanisms. Lastly, considering alternatives to the clustering techniques, such as multiple instance learning and attention-based methods, can provide valuable insights into HR-NMIBC. These approaches enable us to capture the collective behavior of cells within a tissue sample, identify critical regions within whole slide images, and enhance our understanding of the spatial dynamics driving clinical outcomes.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May;71(3):209-249. doi: 10.3322/caac.21660. Epub 2021 Feb 4. PMID: 33538338.

[2] Ma J, Vaishnani DK, Lin R, Lyu J, Ni B, Zhang Y, Hu M, Chen G. Artificial intelligence in bladder cancer: current trends and future possibilities. Chin Med J (Engl). 2022 Apr 5;135(7):881-882. doi: 10.1097/CM9.0000000000001830. PMID: 34985016; PMCID: PMC9276109.

[3] DeGeorge KC, Holt HR, Hodges SC. Bladder Cancer: Diagnosis and Treatment. Am Fam Physician. 2017 Oct 15;96(8):507-514. PMID: 29094888.

[4] Lerner SP, Tangen CM, Sucharew H, Wood D, Crawford ED. Failure to achieve a complete response to induction BCG therapy is associated with increased risk of disease worsening and death in patients with high risk non-muscle invasive bladder cancer. Urol Oncol. 2009 Mar-Apr;27(2):155-9. doi: 10.1016/j.urolonc.2007.11.033. Epub 2008 Mar 4. PMID: 18367117; PMCID: PMC2695968.

[5] Cheng YY, Sun Y, Li J, Liang L, Zou TJ, Qu WX, Jiang YZ, Ren W, Du C, Du SK, Zhao WC. Transurethral endoscopic submucosal en bloc dissection for nonmuscle invasive bladder cancer: A comparison study of HybridKnife-assisted versus conventional dissection technique. J Cancer Res Ther. 2018;14(7):1606-1612. doi: 10.4103/jcrt.JCRT_786_17. PMID: 30589047.

[6] Aldousari S, Kassouf W. Update on the management of non-muscle invasive bladder cancer. Can Urol Assoc J. 2010 Feb;4(1):56-64. doi: 10.5489/cuaj.777. PMID: 20165581; PMCID: PMC2812001.

[7] Xylinas, E., Kent, M., Kluth, L. et al. Accuracy of the EORTC risk tables and of the CUETO scoring model to predict outcomes in non-muscle-invasive urothelial carcinoma of the bladder. Br J Cancer 109, 1460–1466, 2013. https://doi.org/10.1038/bjc.2013.372

[8] Gual Frau J, Palou J, Rodríguez O, Parada R, Breda A, Villavicencio H. Failure of Bacillus Calmette-Guérin therapy in non-muscle-invasive bladder cancer: Definition and treatment options. Arch Esp Urol. 2016 Sep;69(7):423-33. English. PMID: 27617552.

[9] Rossin, G.; Zorzi, F.; Ongaro, L.; Piasentin, A.; Vedovo, F.; Liguori, G.; Zucchi, A.; Simonato, A.; Bartoletti, R.; Trombetta, C.; Pavan, N.; Claps, F. Artificial Intelligence in Bladder Cancer Diagnosis: Current Applications and Future Perspectives. BioMedInformatics 2023, 3, 104-114. https://doi.org/10.3390/biomedinformatics3010008

[10] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective, Artificial Intelligence in Medicine, Volume 23, Issue 1, 2001, Pages 89-109, ISSN 0933-3657, https://doi.org/10.1016/S0933-3657(01)00077-X.

[11] Borhani S, Borhani R, Kajdacsy-Balla A. Artificial intelligence: A promising frontier in bladder cancer diagnosis and outcome prediction. Crit Rev Oncol Hematol. 2022 Mar;171:103601. doi: 10.1016/j.critrevonc.2022.103601. Epub 2022 Jan 19. PMID: 35065220.

[12] Hastie, Trevor, et al. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. 2009 New York: springer.

[13] Klén R, Salminen AP, Mahmoudian M, Syvänen KT, Elo LL, Boström PJ. Prediction of complication related death after radical cystectomy for bladder cancer with machine learning methodology. Scand J Urol. 2019 Oct;53(5):325-331. doi: 10.1080/21681805.2019.1665579. Epub 2019 Sep 25. PMID: 31552774.

[14] Hasnain Z, Mason J, Gill K, Miranda G, Gill IS, Kuhn P, Newton PK. Machine learning models for predicting post-cystectomy recurrence and survival in bladder cancer patients. PLoS One. 2019 Feb 20;14(2):e0210976. doi: 10.1371/journal.pone.0210976. PMID: 30785915; PMCID: PMC6382101.

[15] Cai T, Conti G, Nesi G, Lorenzini M, Mondaini N, Bartoletti R. Artificial intelligence for predicting recurrence-free probability of non-invasive high-grade urothelial bladder cell carcinoma. Oncol Rep. 2007 Oct;18(4):959-64. PMID: 17786360.

[16] Lucas M, Jansen I, van Leeuwen TG, Oddens JR, de Bruin DM, Marquering HA. Deep Learning-based Recurrence Prediction in Patients with Non-muscle-invasive Bladder Cancer. Eur Urol Focus. 2022 Jan;8(1):165-172. doi: 10.1016/j.euf.2020.12.008. Epub 2020 Dec 24. PMID: 33358370.

[17] Qureshi K. N., Naguib R. N.G., Hamdy F. C., Neal D. E. , Mellon J. K., Neural Network Analysis Of Clinicopathological And Molecular Markers In Bladder Cancer, The Journal of Urology, Volume 163, Issue 2, 2000, Pages 630-633, ISSN 0022-5347, https://doi.org/10.1016/S0022-5347(05)67948-7.

[18] Hao S., Xiaoqiang X., Yutao W., Yi L., Chengquan M., Zhigang J., Xiaozhe S., Competitive Risk Model for Specific Mortality Prediction in Patients with Bladder Cancer: A Population-Based Cohort Study with Machine Learning, Journal of Oncology, vol. 2022, Article ID 9577904, 12 pages, 2022. https://doi.org/10.1155/2022/9577904

[19] Ding L, Deng X, Xia W, Wang K, Zhang Y, Zhang Y, Shao X, Wang J. Development and external validation of a novel nomogram model for predicting postoperative recurrence-free survival in non-muscle-invasive bladder cancer. Front Immunol. 2022 Nov 15;13:1070043. doi: 10.3389/fimmu.2022.1070043. PMID: 36458001; PMCID: PMC9706099.

[20] Abuhelwa A.Y., Kichenadasse G., McKinnon, R.A., Rowland, A., Hopkins, A.M., Sorich, M.J., Machine Learning for Prediction of Survival Outcomes with Immune-Checkpoint Inhibitors in Urothelial Cancer. Cancers 2021, 13, 2001. https://doi.org/10.3390/cancers13092001

[21] Bassi P, Sacco E, De Marco V, Aragona M, Volpe A. Prognostic accuracy of an artificial neural network in patients undergoing radical cystectomy for bladder cancer: a comparison with logistic regression analysis. BJU Int. 2007 May;99(5):1007-12. doi: 10.1111/j.1464-410X.2007.06755.x. PMID: 17437435.

[22] Guanjin W., Kin-Man L., Zhaohong D., Kup-Sze C., Prediction of mortality after radical cystectomy for bladder cancer by machine learning techniques, Computers in Biology and Medicine, Volume 63, 2015, Pages 124-132, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2015.05.015

[23] el-Mekresh M, Akl A, Mosbah A, Abdel-Latif M, Abol-Enein H, Ghoneim MA. Prediction of survival after radical cystectomy for invasive bladder carcinoma: risk group stratification, nomograms or artificial neural networks? J Urol. 2009 Aug;182(2):466-72; discussion 472. doi: 10.1016/j.juro.2009.04.018. Epub 2009 Jun 13. PMID: 19524972.

[24] Gavriel CG, Dimitriou N, Brieu N, Nearchou IP, Arandjelović O, Schmidt G, Harrison DJ, Caie PD. Assessment of Immunological Features in Muscle-Invasive Bladder Cancer Prognosis Using Ensemble Learning. Cancers (Basel). 2021 Apr 1;13(7):1624. doi: 10.3390/cancers13071624. PMID: 33915698; PMCID: PMC8036815.

[25] Bhambhvani HP, Zamora A, Shkolyar E, Prado K, Greenberg DR, Kasman AM, Liao J, Shah S, Srinivas S, Skinner EC, Shah JB. Development of robust artificial neural networks for prediction of 5-year survival in bladder cancer. Urol Oncol. 2021 Mar;39(3):193.e7-193.e12. doi: 10.1016/j.urolonc.2020.05.009. Epub 2020 Jun 24. PMID: 32593506.

[26] Ji W, Naguib RN, Ghoneim MA. Neural network-based assessment of prognostic markers and outcome prediction in bilharziasis-associated bladder cancer. IEEE Trans Inf Technol Biomed. 2003 Sep;7(3):218-24. doi: 10.1109/titb.2003.813796. PMID: 14518736.

[27] Lam K. M., He X. J., Choi K. S., Using artificial neural network to predict mortality of radical cystectomy for bladder cancer, International Conference on Smart Computing, Hong Kong, China, 2014, pp. 201-207, doi: 10.1109/SMARTCOMP.2014.7043859.

[28] Guanjin W., Guangquan Z., Kup-Sze C., Kin-Man L., Jie L., Output based transfer learning with least squares support vector machine and its application in bladder cancer prognosis, Neurocomputing, Volume 387, 2020, Pages 279-292, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2019.11.010.

[29] Das, Ashis, Mishra, Shiba, Gopalan, Saji, Survival prediction for bladder cancer using machine learning: development of BlaCaSurv online survival prediction application, 2020 10.1101/2020.11.13.20231191.

[30] Buchner A, May M, Burger M, Bolenz C, Herrmann E, Fritsche HM, Ellinger J, Höfner T, Nuhn P, Gratzke C, Brookman-May S, Melchior S, Peter J, Moritz R, Tilki D, Gilfrich C, Roigas J, Zacharias M, Hohenfellner M, Haferkamp A, Trojan L, Wieland WF, Müller SC, Stief CG, Bastian PJ. Prediction of outcome in patients with urothelial carcinoma of the bladder following radical cystectomy using artificial neural networks. Eur J Surg Oncol. 2013 Apr;39(4):372-9. doi: 10.1016/j.ejso.2013.02.009. PMID: 23465180.

[31] Obajemu, O., Mahfouf, M. and Catto, J.W.F. (2017) A New Fuzzy Modeling Framework for Integrated Risk Prognosis and Therapy of Bladder Cancer Patients. IEEE Transactions on Fuzzy Systems. ISSN 1063-6706

[32] Kolasa, M., Wojtyna, R., Dlugosz, R., Jóźwicki, W. Application of Artificial Neural Network to Predict Survival Time for Patients with Bladder Cancer. In: Kacki, E., Rudnicki, M., Stempczyńska, J. (eds) Computers in Medical Activity. Advances in Intelligent and Soft Computing, vol 65. Springer, Berlin, Heidelberg, 2009, https://doi.org/10.1007/978-3-642-04462-5_11

[33] Eminaga O, Shkolyar E, Breil B, Semjonow A, Boegemann M, Xing L, Tinay I, Liao JC. Artificial Intelligence-Based Prognostic Model for Urologic Cancers: A SEER-Based Study. Cancers. 2022; 14(13):3135. https://doi.org/10.3390/cancers14133135

[34] Song Q, Seigne JD, Schned AR, Kelsey KT, Karagas MR, Hassanpour S. A Machine Learning Approach for Long-Term Prognosis of Bladder Cancer based on Clinical and Molecular Features. AMIA Jt Summits Transl Sci Proc. 2020 May 30;2020:607-616. PMID: 32477683; PMCID: PMC7233061.

[35] Jobczyk M., Stawiski K.,Kaszkowiak M., Rajwa P., Różański W., Soria F., Shariat S. F., Fendler W., Deep Learning-based Recalibration of the CUETO and EORTC Prediction Tools for Recurrence and Progression of Non–muscle-invasive Bladder Cancer, European Urology Oncology, Volume 5, Issue 1, 2022, Pages 109-112, ISSN 2588-9311, https://doi.org/10.1016/j.euo.2021.05.006.

[36] Lucas M, Jansen I, Oddens JR, van Leeuwen TG, Marquering HA, de Bruin DM. Recurrence in non-muscle invasive bladder cancer patients: External validation of the EORTC, CUETO and EAU risk tables and towards a non-linear survival model. Bladder Cancer. 2020;6(3):277-284. doi: 10.3233/BLC-200305

[37] Dovey Z., Pfail J., Martini A., Steineck G., Dey L., Renström L., Hosseini A., Sfakianos J. P., Wiklund P., Bladder Cancer (NMIBC) in a population-based cohort from Stockholm County with long-term follow-up; A comparative analysis of prediction models for recurrence and progression, including external validation of the updated 2021 E.A.U. model, Urologic Oncology: Seminars and Original Investigations, Volume 40, Issue 3, 2022, Pages 106.e1-106.e10, ISSN 1078-1439, https://doi.org/10.1016/j.urolonc.2021.10.008.

[38] Xu X, Wang H, Du P, Zhang F, Li S, Zhang Z, Yuan J, Liang Z, Zhang X, Guo Y, Liu Y, Lu H. A predictive nomogram for individualized recurrence stratification of bladder cancer using multiparametric MRI and clinical risk factors. J Magn Reson Imaging. 2019 Dec;50(6):1893-1904. doi: 10.1002/jmri.26749. Epub 2019 Apr 13. PMID: 30980695; PMCID: PMC6790276.

[39] Catto JW, Linkens DA, Abbod MF, Chen M, Burton JL, Feeley KM, Hamdy FC. Artificial intelligence in predicting bladder cancer outcome: a comparison of neuro-fuzzy modeling and artificial neural networks. Clin Cancer Res. 2003 Sep 15;9(11):4172-7. PMID: 14519642.

[40] López de Maturana E., Picornell A., Masson-Lecomte A. et al., Prediction of non-muscle invasive bladder cancer outcomes assessed by innovative multimarker prognostic models. BMC Cancer 16, 351 (2016). https://doi.org/10.1186/s12885-016-2361-7

[41] Catto JW, Abbod MF, Linkens DA, Larré S, Rosario DJ, Hamdy FC. Neurofuzzy modeling to determine recurrence risk following radical cystectomy for nonmetastatic urothelial carcinoma of the bladder. Clin Cancer Res. 2009 May 1;15(9):3150-5. doi: 10.1158/1078-0432.CCR-08-1960. Epub 2009 Mar 31. PMID: 19336522.

[42] Fujikawa K, Matsui Y, Kobayashi T, Miura K, Oka H, Fukuzawa S, Sasaki M, Takeuchi H, Okabe T. Predicting disease outcome of non-invasive transitional cell carcinoma of the urinary bladder using an artificial neural network model: results of patient follow-up for 15 years or longer. Int J Urol. 2003 Mar;10(3):149-52. doi: 10.1046/j.1442-2042.2003.00589.x. PMID: 12622711.

[43] Catto JW, Abbod MF, Linkens DA, Hamdy FC. Neuro-fuzzy modeling: an accurate and interpretable method for predicting bladder cancer progression. J Urol. 2006 Feb;175(2):474-9. doi: 10.1016/S0022-5347(05)00246-6. PMID: 16406976.

[44] Abbod MF, Linkens DA, Catto JW, Hamdy FC. Comparative study of intelligent models for the prediction of bladder cancer progression. Oncol Rep. 2006;15 Spec no.:1019-22. doi: 10.3892/or.15.4.1019. PMID: 16525693.

[45] Sylvester RJ, Rodríguez O, Hernández V, Turturica D, Bauerová L, Bruins HM, Bründl J, van der Kwast TH, Brisuda A, Rubio-Briones J, Seles M, Hentschel AE, Kusuma VRM, Huebner N, Cotte J, Mertens LS, Volanis D, Cussenot O, Subiela Henríquez JD, de la Peña E, Pisano F, Pešl M, van der Heijden AG, Herdegen S, Zlotta AR, Hacek J, Calatrava A, Mannweiler S, Bosschieter J, Ashabere D, Haitel A, Côté JF, El Sheikh S, Lunelli L, Algaba F, Alemany I, Soria F, Runneboom W, Breyer J, Nieuwenhuijzen JA, Llorente C, Molinaro L, Hulsbergen-van de Kaa CA, Evert M, Kiemeney LALM, N'Dow J, Plass K, Čapoun O, Soukup V, Dominguez-Escrig JL, Cohen D, Palou J, Gontero P, Burger M, Zigeuner R, Mostafid AH, Shariat SF, Rouprêt M, Compérat EM, Babjuk M, van Rhijn BWG. European Association of Urology (EAU) Prognostic Factor Risk Groups for Non-muscle-invasive Bladder Cancer (NMIBC) Incorporating the WHO 2004/2016 and WHO 1973 Classification Systems for Grade: An Update from the EAU NMIBC Guidelines Panel. Eur Urol. 2021 Apr;79(4):480-488. doi: 10.1016/j.eururo.2020.12.033. Epub 2021 Jan 6. Erratum in: Eur Urol. 2023 Feb 23;: PMID: 33419683.

[46] J. . -S. R. Jang, ANFIS: adaptive-network-based fuzzy inference system, IEEE Transactions on Systems, Man, and Cybernetics, vol. 23, no. 3, pp. 665-685, May-June 1993, doi: 10.1109/21.256541.

[47] Schmidt, Uwe et al. "Cell Detection with Star-convex Polygons." International Conference on Medical Image Computing and Computer-Assisted Intervention (2018).

[48] Weigert, Martin & Schmidt, Uwe & Haase, Robert & Sugawara, Ko & Myers. Star-convex Polyhedra for 3D Object Detection and Segmentation in Microscopy, Gene. (2020) 3655-3662. 10.1109/WACV45572.2020.9093435.

[49] Van Gassen, S., Callebaut, B., Van Helden, M.J., Lambrecht, B.N., Demeester, P., Dhaene, T. and Saeys, Y. (2015), FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytometry, 87: 636-645. https://doi.org/10.1002/cyto.a.22625

[50] Wei Shen Tan, Simon Rodney, Benjamin Lamb, Mark Feneley, John Kelly. Management of non-muscle invasive bladder cancer: A comprehensive analysis of guidelines from the United States, Europe and Asia. Cancer Treatment Reviews, Volume 47, 2016, Pages 22-31, ISSN 0305-7372, https://doi.org/10.1016/j.ctrv.2016.05.002.

[51] Chang SS, Boorjian SA, Chou R, Clark PE, Daneshmand S, Konety BR, Pruthi R, Quale DZ, Ritch CR, Seigne JD, Skinner EC, Smith ND, McKiernan JM. Diagnosis and Treatment of Non-Muscle Invasive Bladder Cancer: AUA/SUO Guideline. J Urol. 2016 Oct;196(4):1021-9. doi: 10.1016/j.juro.2016.06.049. Epub 2016 Jun 16. PMID: 27317986.

[52] T. Takagi, M. Sugeno, Derivation of Fuzzy Control Rules from Human Operator's Control Actions, IFAC Proceedings Volumes, Volume 16, Issue 13, 1983, Pages 55-60, ISSN 1474-6670, https://doi.org/10.1016/S1474-6670(17)62005-6.

[53] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (COLT '92). Association for Computing Machinery, New York, NY, USA, 144–152. https://doi.org/10.1145/130385.130401

[54] Rokach, Lior & Maimon, Oded. (2005). Decision Trees. 10.1007/0-387-25465-X_9.

[55] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.

[56] L. Breiman, "Bagging predictors", Machine Learning, 24(2), 123-140, 1996.

[57] L. Breiman, "BIAS, VARIANCE, AND ARCING CLASSIFIERS" (PDF). TECHNICAL REPORT, 1996

[58] L. Breiman, "Random Forests", Machine Learning, 45(1), 5-32, 2001.

[59] Fuster, S., Khoraminia, F., Eftestøl, T., Zuiverloon, T. and Engan, K., 2023. Active Learning Based Domain Adaptation for Tissue Segmentation of Histopathological Images. arXiv preprint arXiv:2303.05225.

[60] van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T, the scikit-image contributors. 2014. scikit-image: image processing in Python. PeerJ 2:e453 https://doi.org/10.7717/peerj.453

[61] Slotman A, Xu M, Lindale K, Hardy C, Winkowski D, Baird R, Chen L, Lal P, van der Kwast T, Jackson CL, Gooding RJ, Berman DM. Quantitative Nuclear Grading: An Objective, Artificial Intelligence-Facilitated Foundation for Grading Noninvasive Papillary Urothelial Carcinoma. Lab Invest. 2023 Apr 13;103(7):100155. doi: 10.1016/j.labinv.2023.100155. Epub ahead of print. PMID: 37059267.

[62] Azhari, Mourad & Alaoui, Altaf & Acharoui, Zakia & Ettaki, Badia & Zerouaoui, Jamal. (2019). Adaptation of the random forest method: solving the problem of pulsar search. SCA '19: Proceedings of the 4th International Conference on Smart City Applications. 1-6. 10.1145/3368756.3369004.

[63] Jobczyk M, Stawiski K, Fendler W, Różański W. Validation of EORTC, CUETO, and EAU risk stratification in prediction of recurrence, progression, and death of patients with initially non-muscle-invasive bladder cancer (NMIBC): A cohort analysis. Cancer Med. 2020 Jun;9(11):4014-4025. doi: 10.1002/cam4.3007. Epub 2020 Mar 26. PMID: 32216043; PMCID: PMC7286464.

A

Appendix A

Clinical Parameter	Subgroup	All	BCG treatment		Progression		HG recurrence	
			Responders	Failure	Yes	No	Yes	No
ALL		720	525 (73%)	195 (27%)	100 (14%)	620 (86%)	110 (15%)	610 (85%)
Gender	Male	575 (80%)	418 (72%)	160 (28%)	88 (15%)	490 (85%)	95 (16%)	483 (84%)
	Female	145 (20%)	107 (77%)	35 (23%)	12 (12%)	130 (88%)	15 (14%)	127 (86%)
Age (years)	Median (min-max)	71 (31-98)	72 (32-98)	73 (45-98)	72.5 (41-92)	72 (31-98)	73 (45-92)	72 (32-98)
Smoking	Yes	430 (60%)	320 (73%)	120 (27%)	60 (14%)	381 (86%)	63 (11%)	367 (89%)
	No	290 (40%)	206 (71%)	84 (29%)	45 (16%)	245 (84%)	48 (17%)	242 (83%)
Size (cm)	≤ 3	665 (92%)	485 (73%)	180 (27%)	97 (15%)	568 (85%)	107 (16%)	558 (84%)
	> 3	54 (8%)	41 (76%)	14 (24%)	8 (15%)	46 (85%)	8 (15%)	46 (85%)
Staging	Tis	77 (11%)	56 (73%)	21 (27%)	10 (13%)	67 (87%)	13 (17%)	64 (83%)
	Ta	253 (35%)	204 (81%)	39 (19%)	36 (14%)	217 (86%)	29 (12%)	224 (88%)
	T1	389 (54%)	267 (69%)	122 (31%)	60 (15%)	329 (85%)	73 (19%)	316 (81%)
Grading	G1L	12 (2%)	11 (92%)	1 (8%)	0 (0%)	12 (100%)	1 (7%)	11 (93%)
	G2L	32 (4%)	25 (80%)	8 (20%)	8 (23%)	24 (77%)	2 (8%)	29 (92%)
	G2H	46 (6%)	36 (79%)	10 (21%)	10 (22%)	36 (78%)	4 (9%)	42 (91%)
	G3H	629 (87%)	454 (72%)	175 (18%)	88 (14%)	541 (86%)	108 (17%)	521 (83%)
Concomitant CIS	Yes	37 (5%)	24 (66%)	13 (34%)	17 (45%)	20 (55%)	5 (15%)	32 (85%)
	No	682 (95%)	502 (74%)	180 (26%)	89 (13%)	593 (87%)	109 (16%)	573 (84%)
History of cancer	Yes	161 (22%)	125 (78%)	36 (22%)	21 (13%)	140 (87%)	23 (14%)	138 (86%)
	No	558 (77%)	402 (72%)	156 (28%)	84 (15%)	474 (85%)	92 (16%)	466 (84%)
Number of tumors	Single	305 (42%)	235 (75%)	70 (25%)	28 (17%)	277 (83%)	41 (14%)	264 (86%)
	Multiple	224 (31%)	156 (70%)	68 (3%)	25 (11%)	199 (89%)	46 (21%)	178 (79%)
Follow-up (months)	Median (min-max)	60 (2-215)	72 (2-210)	60 (7-215)	48 (7-180)	72 (2-215)	48 (3-215)	72 (2-215)

Table A.1: Training set overview

RECURRENCE

Table A.2: 6 months recurrence

Study	Study aim (data set size)	Features	Model	Performance	Findings
Qureshi et al. (2000) [17]	6 months recur- rence (212)	Stage, grade, tumor size, tumor number, gender, and EGFR status, smoking habit, histology, cis presence, metapla- sia, architecture, site, c-erbB2, and p53 status.	ANN implemented with NeuralWorks Professional II/Plus software	Se: 0.7, Sp: 0.8, Ac: 0.75	Tumor size shown to be the most important feature

Table A.3: 1 year recurrence

Study	Study aim (data set size)	Features	Model	Performance	Findings
Hasnain et al. (2019) [12]	1 year after RC (3499)	Operative findings at transurethral resection and radical cystectomy, pathology	Meta classifier based on SVM, bagged SVM, KNN, AdaBoost, RF and GBT	Se: 0.739, Sp: 0.714, Ac: 0.388	Use of the information theory concept of mutual information (MI) to uncover correlated parameters
Lucas et al. (2020) [36]	1 year re-currence (452)	WHO'73 grading, the number of tumors as defined by the CUETO, the recurrence rate as defined by the EORTC and the age classification as defined by the CUETO	Cox proportional hazards, Boosted Cox model and Random survival forest	Se: 0.73 (CPH), 0.70 (BCM), 0.71 (RSF); Sp: 0.59 (CPH), 0.58 (BCM), 0.53 (RSF); Ac: 0.60 (CPH), 0.59(BCM), 0.55(RSF); AUC: 0.66 (CPH), 0.7 (BCM), 0.62 (RSF); C-index: 0.61 (overall CPH), 0.64 (overall BCM), 0.61 (overall RSF)	EORTC and EAU risk classification show slightly better predictive value than the CUETO risk stratification in the population under study. The study also highlights the subjectivity in assessing the histopathological variables and the difficulty in assessing grading, staging, and muscular is propria. The study suggests that new prognostic markers are needed
Dovey et al. (2022) [37]	1 year re-currence (395)	Multifocality, tumor stage, grade and size	Simplified model based on EORTC and CUETO	AUC: 0.7	EORTC, CUETO and EAU Sylvester et al. (2021) WHO '73 and '04/16 models tend to underestimate recurrence. It is suggested that the use of clinical covariables to predict recurrence may have reached their upper limits of accuracy and studies have investigated molecular subtyping and genomic classification as an alternative

Table A.4: 2 years recurrence

Study	Study aim (data set size)	Features	Model	Performance	Findings
Dovey et al. (2022) [37]	2 years recur-rence (395)	Multifocality, tumor stage, grade and size	Simplified model based on EORTC and CUETO	AUC: 0.67	EORTC, CUETO and EAU Sylvester et al. (2021) WHO ‘73 and ‘04/16 models tend to underestimate recurrence. It is suggested that the use of clinical covariables to predict recurrence may have reached their upper limits of accuracy and studies have investigated molecular subtyping and genomic classification as an alternative
Xu et al. (2019) [38]	2 years risk of re-currence (71)	Age at the time of initial surgery, gender, histological grade, MIS of the archived tumor with the maximal size in bladder lumen, tumor size, NoT, and operation choice (TURBT or RC)	Nomogram based on the features selected with a Rad_Score model constructed using support vector machine-based recursive feature elimination (SVM-RFE) approach and logistic regression	Se: 0.778; Sp: 0.738; Ac: 0.755; AUC: 0.822	The most promising features extracted were image-related

Table A.5: 3 years recurrence

Study	Study aim (data set size)	Features	Model	Performance	Findings
Hasnain et al. (2019) [12]	3 years after RC (3499)	Operative findings at transurethral resection and radical cystectomy, pathology	Meta classifier based on SVM, bagged SVM, KNN, AdaBoost, RF and GBT	Se: 0.72, Sp: 0.708, Ac: 0.535	Use of the information theory concept of mutual information (MI) to uncover correlated parameters

Table A.6: 5 years recurrence

Study	Study aim (data set size)	Features	Model	Performance	Findings
Hasnain et al. (2019) [12]	5 years after RC (3499)	Operative findings at transurethral resection and radical cystectomy, pathology	Meta classifier based on SVM, bagged SVM, KNN, AdaBoost, RF and GBT	Se: 0.7, Sp: 0.702, Ac: 0.588	pathologic stage subgroup, pT stage, pN stage, pM stage, number of positive lymph nodes, pathologic positive lymph nodes, pathologic lymphovascular invasion, clinical T stage (preoperative) shown to be the most important features
Buchner et al. (2013) [30]	5 years recurrence (2111)	Age, gender, tumour stage and grade (in transurethral resection of the bladder/TURB and RC), carcinoma in situ (TURB and RC), pathological lymph node status and lymphovascular invasion	ANN with a three-layer feed-forward perceptron architecture	Se: 0.4, Sp: 0.895, Ac: 0.74	Lymphovascular invasion, pathological T stage and pathological lymph node status shown to be the most important features
Lucas et al. (2020) [36]	5 years recurrence (452)	WHO'73 grading, the number of tumors as defined by the CUETO, the recurrence rate as defined by the EORTC and the age classification as defined by the CUETO	Cox proportional hazards, Boosted Cox model and Random survival forest	Se: 0.58 (CPH), 0.64 (BCM), 0.59 (RSF); Sp: 0.62 (CPH), 0.61 (BCM), 0.65 (RSF); Ac: 0.60 (CPH), 0.60(BCM), 0.61(RSF); AUC: 0.69 (CPH), 0.72 (BCM), 0.69 (RSF); C-index: 0.61 (overall CPH), 0.64 (overall BCM), 0.61 (overall RSF)	EORTC and EAU risk classification show slightly better predictive value than the CUETO risk stratification in the population under study. The study also highlights the subjectivity in assessing the histopathological variables and the difficulty in assessing grading, staging, and muscular is propria. The study suggests that new prognostic markers are needed
Dovey et al. (2022) [37]	2 years recurrence (395)	Multifocality, tumor stage, grade and size	Simplified model based on EORTC and CUETO	AUC: 0.69	EORTC, CUETO and EAU Sylvester et al. (2021) WHO '73 and '04/16 models tend to underestimate recurrence. It is suggested that the use of clinical covariables to predict recurrence may have reached their upper limits of accuracy and studies have investigated molecular subtyping and genomic classification as an alternative

Table A.7: 80 months recurrence

Study	Study aim (data set size)	Features	Model	Performance	Findings
Catto et al. (2003) [39]	80 months recurrence after surgery (109)	Stage, grade, age, sex, smoking exposure, previous cancers	Neuro Fuzzy Modelling	Se: 0.92, Sp: 0.9, Ac: 0.92, AUC: 0.98	Tumor grade, patient age, smoking history, and p53 expression shown to be the most important features

Table A.8: 10 years recurrence

Study	Study aim (data set size)	Features	Model	Performance	Findings
Dovey et al. (2022) [37]	10 years recur-rence (395)	Multifocality, tumor stage, grade and size	Simplified model based on EORTC and CUETO	AUC: 0.66	EORTC, CUETO and EAU Sylvester et al. (2021) WHO '73 and '04/16 models tend to underestimate recurrence. It is suggested that the use of clinical covariables to predict recurrence may have reached their upper limits of accuracy and studies have investigated molecular subtyping and genomic classification as an alternative

Table A.9: 15 years recurrence

Study	Study aim (data set size)	Features	Model	Performance	Findings
Fujikawa et al. (2002) [42]	15 years recur-rence (90)	Tumor stage, grade,tumor number, age, gender, tumor architecture and estimates of mean nuclear volume	Bayesian neural tool of SPSS Neural Connection 2.1 software	Se: 0.33, Sp: 0.4	The ANN model could not predict tumor recurrence with the features included in this study

Table A.10: Time to first recurrence

Study	Study aim (data set size)	Features	Model	Performance	Findings
López de Matu-rana et a. (2016) [40]	Time to first recurrence (1105)	Area, gender, number of tumours, tumour stage and grade, tumour size, treatment	Sequential threshold model	AUC: 0.62	Role of common SNPs is very limited in the prediction of risk of recurrence and future studies should explore whether the integration of other genetic variants. Time-to-first-recurrence (TFR), defined as the reappearance of a NMIBC tumor following a previous negative follow-up cystoscopy
Catto et al. (2009) [41]	Time to first recurrence up to 20 years after RC and PLND (609)	Gender, pathologic stage, pathologic grade, carcinoma in situ and lymphovascular invasion	Two Neuro Fuzzy Modelling combined in series to predict the risk and the timing of post-RC tumor recurrence	Se: 0.81, Sp: 0.85, C-index: 0.92	Tumor stage, lymphovascular invasion and the number of removed lymph nodes shown to be the most important features

PROGRESSION

Table A.11: 6 months progression

Study	Study aim (data set size)	Features	Model	Performance	Findings
Qureshi et al. (2000) [17]	6 months progression (212)	Stage, grade, tumor size, tumor number, gender, and EGFR status	ANN implemented with NeuralWorks Professional II/Plus software	Se: 0.7, Sp: 0.82, Ac: 0.8	Epidermal growth factor receptor (EGFR) shown to be the most important feature

Table A.12: 1 year progression

Study	Study aim (data set size)	Features	Model	Performance	Findings
Dovey et al. (2022) [37]	1 year progres- sion (395)	Multifocality, tumor stage, grade and size	Simplified model based on EORTC and CUETO	AUC: 0.88	The main limitation of this study is the relatively large proportion of low risk NMIBC patients, resulting in smaller numbers for the analysis of progression data. It is suggested that the use of clinical covariables to predict recurrence may have reached their upper limits of accuracy and studies have investigated molecular subtyping and genomic classification as an alternative

Table A.13: 2 years progression

Study	Study aim (data set size)	Features	Model	Performance	Findings
Dovey et al. (2022) [37]	1 year progres- sion (395)	Multifocality, tumor stage, grade and size	Simplified model based on EORTC and CUETO	AUC: 0.88	The main limitation of this study is the relatively large proportion of low risk NMIBC patients, resulting in smaller numbers for the anal- ysis of progression data. It is suggested that the use of clinical covariables to predict recurrence may have reached their upper limits of accuracy and studies have investigated molecular subtyping and genomic classification as an alternative

Table A.14: 5 years progression

Study	Study aim (data set size)	Features	Model	Performance	Findings
Dovey et al. (2022) [37]	1 year progres- sion (395)	Multifocality, tumor stage, grade and size	Simplified model based on EORTC and CUETO	AUC: 0.84	The main limitation of this study is the relatively large proportion of low risk NMIBC patients, resulting in smaller numbers for the anal- ysis of progression data. It is suggested that the use of clinical covariables to predict recurrence may have reached their upper limits of accuracy and studies have investigated molecular subtyping and genomic classification as an alternative

Table A.15: 10 years progression

Study	Study aim (data set size)	Features	Model	Performance	Findings
Dovey et al. (2022) [37]	1 year progres- sion (395)	Multifocality, tumor stage, grade and size	Simplified model based on EORTC and CUETO	AUC: 0.82	The main limiation of this study is the relatively large proportion of low risk NMIBC patients, resulting in smaller numbers for the anal- ysis of progression data. It is suggested that the use of clinical covariables to predict recurrence may have reached their upper limits of accuracy and studies have investigated molecular subtyping and genomic classification as an alternative

Table A.16: 80 months progression

Study	Study aim (data set size)	Features	Model	Performance	Findings
Catto et al. (2005) [43]	80 months progres- sion (117)	Tumor stage, tumor grade, age, gender, smoking exposure, previous cancers	Hybrid neural network that combines a fuzzy logic model trained on subgroups defined by hierarchical clustering algorithm	Se: 0.88, Sp: 0.99, Ac: 0.94, AUC: 0.99	Age, grade, stage, smoking and methylation shown to be the most important feature

Table A.17: 100 months progression

Study	Study aim (data set size)	Features	Model	Performance	Findings
Abbod et al. (2005) [44]	100 months progres-sion (117)	Stage, grade, age, sex, smoking exposure, previous cancers	Neuro-fuzzy modelling (NFM) and a multi-layered perceptron artificial neural networks (ANN) with 15 hidden neurones	Se: 0.81 (ANN), 0.88(NFM); Sp: 0.95(ANN), 0.99 (NFM); Ac: 0.89 (ANN), 0.94 (NFM)	Smoking was significantly related to more advanced disease compared to non-smoking

Table A.18: 15 years progression

Study	Study aim (data set size)	Features	Model	Performance	Findings
Fujikawa et al. (2002) [42]	15 years progres-sion (90)	Tumor stage, grade,tumor number, age, gender, tumor architecture and estimates of mean nuclear volume	Bayesian neural tool of SPSS Neural Connection 2.1 software	Se: 1, Sp: 0.67	Patients who were judged to have a favorable prognosis using ANN analysis did not progress within the 15-year follow-up period

Table A.19: Time to first progression

Study	Study aim (data set size)	Features	Model	Performance	Findings
López de Maturana et a. (2016) [40]	Time to first pro-gression (1105)	Area, age, number of tumours, tumour stage and grade, number of recurrences and treatment	Sequential thresh-old model	AUC: 0.76	Role of common SNPs is very limited in the prediction of risk of recurrence and future studies should explore whether the integration of other genetic variants. Time to progression defined as the development of a muscle invasive tumor or a metastatic disease, or death because of UCB, after a previous diagnosis of NMIBC

B

Appendix B

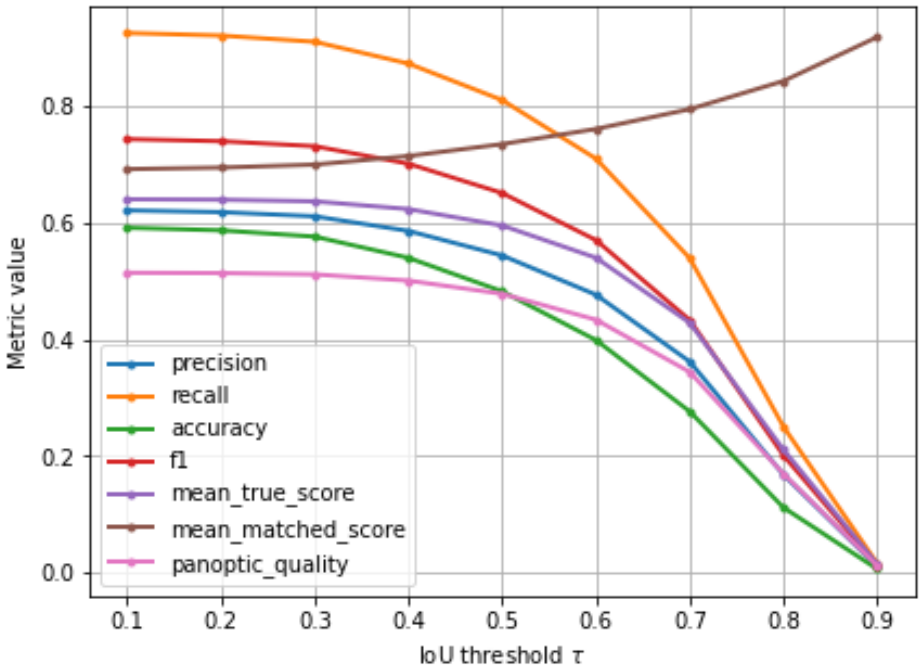

Figure B.1: Pretrained segmentation model

Clinicopathological data (n=504)

Table B.1: BCG failure

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
NFM	Clinical data (n=504)	0.57	0.60	0.57	0.72	0.42
RF	Clinical data (n=504)	0.70	0.72	0.70	0.77	0.63
GB	Clinical data (n=504)	0.58	0.60	0.58	0.63	0.53
SVM	Clinical data (n=504)	0.58	0.62	0.58	0.67	0.50
DT	Clinical data (n=504)	0.57	0.59	0.57	0.63	0.50
ET	Clinical data (n=504)	0.62	0.65	0.62	0.70	0.53

Table B.2: Progression

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
NFM	Clinical data (n=504)	0.63	0.63	0.63	0.64	0.62
RF	Clinical data (n=504)	0.68	0.69	0.70	0.73	0.67
GB	Clinical data (n=504)	0.71	0.73	0.71	0.72	0.72
SVM	Clinical data (n=504)	0.67	0.67	0.67	0.67	0.67
DT	Clinical data (n=504)	0.64	0.67	0.64	0.72	0.56
ET	Clinical data (n=504)	0.64	0.65	0.64	0.67	0.61

Table B.3: HG recurrence

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
NFM	Clinical data (n=504)	0.55	0.51	0.55	0.58	0.52
RF	Clinical data (n=504)	0.61	0.62	0.61	0.63	0.58
GB	Clinical data (n=504)	0.63	0.63	0.63	0.63	0.63
SVM	Clinical data (n=504)	0.55	0.54	0.55	0.53	0.58
DT	Clinical data (n=504)	0.61	0.63	0.61	0.68	0.53
ET	Clinical data (n=504)	0.37	0.37	0.37	0.37	0.37

Image features

Table B.4: BCG failure

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
NFM	Image features (n=504)	0.59	0.62	0.60	0.72	0.45
RF	Image features (n=504)	0.67	0.70	0.67	0.75	0.64
GB	Image features (n=504)	0.60	0.61	0.60	0.62	0.55
SVM	Image features (n=504)	0.59	0.61	0.59	0.68	0.53
DT	Image features (n=504)	0.55	0.56	0.55	0.62	0.53
ET	Image features (n=504)	0.64	0.67	0.64	0.69	0.57

Table B.5: Progression

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
NFM	Image features (n=504)	0.65	0.67	0.65	0.64	0.62
RF	Image features (n=504)	0.70	0.71	0.70	0.74	0.67
GB	Image features (n=504)	0.69	0.70	0.69	0.70	0.72
SVM	Image features (n=504)	0.68	0.67	0.68	0.65	0.67
DT	Image features (n=504)	0.63	0.64	0.64	0.70	0.59
ET	Image features (n=504)	0.61	0.62	0.61	0.69	0.64

Table B.6: HG recurrence

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
NFM	Image features (n=504)	0.58	0.59	0.58	0.56	0.53
RF	Image features (n=504)	0.64	0.62	0.64	0.62	0.60
GB	Image features (n=504)	0.62	0.63	0.62	0.61	0.61
SVM	Image features (n=504)	0.58	0.56	0.58	0.55	0.59
DT	Image features (n=504)	0.60	0.61	0.60	0.66	0.54
ET	Image features (n=504)	0.45	0.48	0.45	0.51	0.49

Clinicopathological data and image features

Table B.7: BCG failure

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
NFM	Clinical data and image features (n=504)	0.59	0.57	0.58	0.59	0.58
RF	Clinical data and image features (n=504)	0.77	0.77	0.77	0.80	0.73
GB	Clinical data and image features (n=504)	0.73	0.75	0.73	0.80	0.67
SVM	Clinical data and image features (n=504)	0.62	0.63	0.62	0.67	0.57
DT	Clinical data and image features (n=504)	0.57	0.59	0.57	0.63	0.50
ET	Clinical data and image features (n=504)	0.65	0.68	0.65	0.73	0.57

Table B.8: Progression

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
NFM	Clinical data and image features (n=504)	0.59	0.57	0.58	0.59	0.58
RF	Clinical data and image features (n=504)	0.78	0.80	0.78	0.84	0.72
GB	Clinical data and image features (n=504)	0.73	0.75	0.73	0.74	0.77
SVM	Clinical data and image features (n=504)	0.64	0.67	0.64	0.72	0.56
DT	Clinical data and image features (n=504)	0.72	0.72	0.72	0.72	0.72
ET	Clinical data and image features (n=504)	0.69	0.70	0.69	0.72	0.67

Table B.9: HG recurrence

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
NFM	Clinical data and image features (n=504)	0.59	0.57	0.58	0.59	0.58
RF	Clinical data and image features (n=504)	0.71	0.72	0.71	0.74	0.68
GB	Clinical data and image features (n=504)	0.66	0.65	0.66	0.63	0.68
SVM	Clinical data and image features (n=504)	0.63	0.63	0.63	0.63	0.63
DT	Clinical data and image features (n=504)	0.58	0.58	0.58	0.58	0.58
ET	Clinical data and image features (n=504)	0.53	0.50	0.53	0.47	0.58

Clinicopathological data (n=900)

Table B.10: BCG failure

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
NFM	Clinical data (n=900)	0.57	0.60	0.57	0.72	0.42
RF	Clinical data (n=900)	0.66	0.68	0.66	0.71	0.61
GB	Clinical data (n=900)	0.61	0.65	0.61	0.70	0.52
SVM	Clinical data (n=900)	0.52	0.53	0.52	0.53	0.52
DT	Clinical data (n=900)	0.61	0.62	0.61	0.61	0.61
ET	Clinical data (n=900)	0.60	0.61	0.60	0.64	0.56

Table B.11: Progression

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
NFM	Clinical data (n=900)	0.63	0.63	0.63	0.64	0.62
RF	Clinical data (n=900)	0.77	0.76	0.77	0.75	0.74
GB	Clinical data (n=900)	0.75	0.76	0.75	0.77	0.72
SVM	Clinical data (n=900)	0.52	0.52	0.52	0.52	0.52
DT	Clinical data (n=900)	0.63	0.63	0.63	0.65	0.61
ET	Clinical data (n=900)	0.60	0.63	0.60	0.68	0.52

Table B.12: HG recurrence

Model	Method	Accuracy	F1 score	AUC	Sensitivity	Specificity
NFM	Clinical data (n=900)	0.55	0.51	0.55	0.58	0.52
RF	Clinical data (n=900)	0.63	0.66	0.63	0.71	0.54
GB	Clinical data (n=900)	0.74	0.74	0.74	0.72	0.76
SVM	Clinical data (n=900)	0.49	0.54	0.49	0.60	0.55
DT	Clinical data (n=900)	0.63	0.66	0.63	0.71	0.54
ET	Clinical data (n=900)	0.60	0.63	0.60	0.69	0.51